REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1/28/2005	Phase I Final Report 2/1/04 - 12/31/04

4. TITLE AND SUBTITLE
A Joint Feature Extraction and Data Compression Method
For Low Bit Rate Transmission In Distributed Acoustic Sensor Environments

5. FUNDING NUMBERS
W15QKN-04-C-1061

6. AUTHOR(S)
M.R. Azimi-Sadjadi and A. Pezeshki

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Information System Technologies, Inc.
5412 Hilldale Court
Fort Collins, CO 80526

8. PERFORMING ORGANIZATION REPORT NUMBER
IST0006

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
US Army TACOM-ARDEC
AMSTA-AR-FSF-R, Bldg 407
Picatinny Arsenal, NJ 07806-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Report developed under SBIR contract for topic A03-001
Unattended distributed passive acoustic sensors are widely used for remote battlefield surveillance, situation awareness and monitoring applications. To improve the spatial resolution for separating multiple closely spaced targets while reducing the on-board computational requirements, a modest quantity of single microphones could be deployed in a surveillance area of interest. These distributed microphones are considerably less expensive, small sized and contain generic DSP boards capable of performing detection, feature extraction and data compression tasks. They are equipped with basic communication systems to transmit essential compressed target information to a master station, which has more computational power to carry out high-level operations for sensor array processing and target classification. In this Phase I research, a subband-based joint detection, feature extraction, data compression/encoding system for low bit rate transmission of essential target information will be developed. The extracted features allow for detection and classification of the targets as well as data compression/encoding without incurring degradation in the overall performance. New methods for formation of the optimal sparse sensor arrays based upon multi-channel coherence information would also be developed. The effectiveness of the developed methods will be demonstrated on real and synthesized data sets.

14. SUBJECT TERMS
SBIR Report

15. NUMBER OF PAGES
41

16. PRICE CODE

17. SECURITY CLASSIFICTION OF REPORT	18. SECURITY CLASSIFICTION OF THIS PAGE	19. SECURITY CLASSIFICTION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

Final Report

SBIR-Army Phase I

A Joint Feature Extraction and Data Compression Method For Low Bit Rate Transmission In Distributed Acoustic Sensor Environments

PI: Dr. M. R. Azimi-Sadjadi

CO-PI Dr. A. Pezeshki

Information System Technologies, Inc.
425 W. Mulberry St., Suite 108
Fort Collins, CO 80521
Tel: (970) 224-2556
E-mail: mo@infsyst.biz
website: www.infsyst.com

Submitted to :

Mr. Robert Wade
US Army TACOM-ARDEC
AMSTA-AR-FSF-R, Bldg 407
Picatinny Arsenal, NJ 07806-5000
Tel : (973) 724-5356, E-mail: wade@pica.army.mil

Contract No.: W15QKN-04-C-1061

Duration of Effort: February 1, 2004 - December 31, 2004

Contents

1 Introduction and Objectives of Phase I Research **1**

2 Identification of Barrier Issues & Review of Existing Methods **3**
- 2.1 Target Attribute Detection . . . 3
- 2.2 Time-Frequency Features & Data Compression . . . 4
- 2.3 Encoding Methods . . . 5

3 A joint Subband Detection, Feature Extraction and Data Compression System **8**
- 3.1 Subband Target Attribute Detection . . . 8
- 3.2 Subband Feature Extraction, Encoding and Decoding Processes . . . 10
- 3.3 High-Resolution Wideband DOA Estimation Method . . . 11
  - 3.3.1 Geometric Wideband DOA Estimation . . . 12
- 3.4 Vehicle Classification Using Subband Features . . . 13
- 3.5 Optimum Array Formation and Coherence Analysis . . . 13
  - 3.5.1 Multiple-Channel Coherence Analysis . . . 14

4 Test Results & Observations **14**
- 4.1 Brief Description of the Acquired Data Sets . . . 14
- 4.2 Subband Detection, Feature Extraction and Data Compression Results . . . 17
  - 4.2.1 Study 1: Optimum Number of Subbands at a Fixed Bit Budget . . . 17
  - 4.2.2 Study 2: Optimum Bit Rate for Fixed Number of Subbands . . . 18
  - 4.2.3 Study 3: Optimum Subband and Optimum Bit Rate . . . 22
  - 4.2.4 Study 4: Comparison with Wavelet-Based Method . . . 26
- 4.3 Subband Vehicle Classification Results . . . 28
- 4.4 Coherence Analysis . . . 30

5 Conclusions and Suggestions for Future Work **35**

1 Introduction and Objectives of Phase I Research

Unattended passive acoustic sensors [1]-[3] are among the widely used sensors for remote battlefield surveillance, situation awareness and monitoring applications. Together with dedicated digital signal processing (DSP), these small and cost effective sensors can provide real-time information about different types of ground and airborne targets. They are rugged and reliable and can be left in the field for a relatively long period of time after deployment. The DSP associated with these sensors must be capable of performing various tasks including target detection and direction of arrival (DOA) estimation, tracking, target classification and identification. Generally, there can be a wide variety of target types in battlefield depending on the specific mission, e.g. ground targets (trucks, tanks, etc), airborne targets (helicopters, missiles, airplanes), or personnel in urban areas. These sources typically have signatures that overlap both temporally and spectrally. In addition, optimum performance is highly dependent on terrain, weather, and background noise.

The work at the Army Research Laboratory (ARL) [2],[3] involved development of several array processing algorithms using baseline acoustic arrays to estimate DOA's of moving targets from their acoustic signatures. The experimental results in these references, using a circular array of six sensors with a radius of 4 ft, indicated the promise of the algorithms for acoustic source tracking. However, the existing algorithms have limited capability when applied to realistic multiple target scenarios especially in adverse operating and environmental conditions. Information System Technologies, Inc (ISTI) has developed, as part of Army SBIR Phase II efforts [4], various high resolution DOA estimation and blind source separation algorithms for resolving several closely spaced targets in a group. The effectiveness of these methods have been tested on both synthesized and real SAFE II data.

Although the idea of deploying several baseline circular microphone arrays is found to be effective in detecting and tracking several isolated groups of targets, it cannot typically resolve the targets within a group when they are spatially very close together [4]. This is due to limited spatial sampling resolution. Additionally, the number of target groups or individual targets that can be resolved using these baseline arrays is typically limited by the number of sensors in the array. Furthermore, these systems are not cost effective owing to the high computational requirements for their bilateral communication with a master station, local processing needed for detection and DOA estimation and possibly the associated global positioning systems (GPS) to determine their locations in the field.

Our recent studies [4] on coherence analysis of real SAFE II data using canonical coordinates decomposition [5] revealed the fact that in nominal operating conditions the signals collected at different sensory nodes are indeed coherent even at distant ($\sim$ 500m) node separation. It was observed that coherence was only lost in extremely windy conditions. This study suggested the possibility of exploring an entirely different approach to this problem that would entail distributing a modest quantity of single microphones in a surveillance area of interest (e.g. $1km^2$). These microphones are considerably less expensive, small sized and contain generic off-the-shelf DSP boards capable of performing simple detection, feature extraction and data compression tasks. They are also equipped with basic communication systems sufficient to transmit essential compressed target information to a master station which has more sophisticated computational capabilities to receive, process the data from all or a subset of the distributed sensors and perform high-level operations such as DOA estimation, data association, tracking, and target classification. This new paradigm offers numerous expected benefits comparing to the conventional baseline array deployment scheme. These include: better spatial resolution in dealing with multiple closely spaced targets within a group, ability to detect larger number of targets, less hardware complexity and hence considerably lower costs, more flexibility in configuring different arrays based upon particular scenario by using subsets or clusters of microphones. The latter implies that the configuration, resolution, and tracking ability of the

arrays are greatly adaptable to a particular military scenario that involves multiple groups or formations of several closely spaced targets.

However, there are several key issues of paramount importance in this new paradigm that must be carefully addressed before actual implementation. One of the main issues is the development of a simple and yet robust detection scheme [6] to determine the existence of target indication based upon some clues (features) in the acoustic signatures. This is extremely important for any real target tracking environment since there is usually no *a priori* knowledge as to when the targets initiate and terminate in the surveillance area of the sensors. Thus, the system must operate continuously and screen the data for presence of targets at certain number of measurements as soon as there is sufficient information for making high confidence decision. Clearly, the major benefit of this process is that there is no need to transmit the sensor output unless it exhibits convincing evidence of target presence. This provides significant overall data reduction that is ideally needed for efficient processing and communication of the relevant information to a master station. Once target indication is detected in the recorded signals of all or a subset of the sensors, the corresponding subbands that carry useful target information must then be compressed and encoded [7]-[9]. The idea is to extract essential target attributes and hence remove redundancy in the data to be transmitted. After redundancy removal, the signals will be encoded using substantially lesser number of bits and hence can be transmitted more efficiently. For our particular problem, the temporal-spectral characteristics of different types of targets are exploited in order to extract a low-dimensional spectral feature set that carries enough information to allow for accurate DOA estimation and target classification. To reduce the computational requirements of the overall process it is essential to combine the detection, feature extraction, data compression, and classification processes into a joint process with various functional capabilities.

The previous work [10]-[19] in the area of acoustic, sonar, and speech data compression typically use transform-based methods such as discrete cosine transform (DCT) or variants of the wavelet transform. A detailed review of these methods is provided in Section 2. However, to the best of our knowledge, none of the existing data compression methods can directly be applied to acoustic signatures of multiple targets recorded by several distributed microphones to achieve a compression ratio of at least 50:1. There are several critical requirements and issues that should be considered in developing a joint feature extraction-data compression scheme for this particular problem. These are: preservation of time delay information of the sources for successful DOA estimation and tracking at the master computer, preservation of essential target signature features for target classification, compliance with bit budget or compression ratio requirements for each sensor, optimal array formation from a subset of microphones, and global synchronization of the sensors in an array.

The goal of the Phase I research effort is to investigate and develop new and efficient joint feature extraction and data compression algorithms that achieve low bit rate transmission of less than 1kbps for the existing air-deployed acoustic sensor (e.g. OMNI-400 series [1]), while incurring minimal degradation in DOA estimation accuracy as compared to the performance on the uncompressed data. In addition, it is important that essential spectral and tonal features of the targets are kept for source separation and classification. The effectiveness of the developed algorithms is demonstrated on multiple target acoustic signatures acquired from the US Army-TACOM-ARDEC, Picatinny Arsenal,NJ.

The organization of this final report is as follows. Section 2 provides a review of the existing methods for target attributes detection, time-frequency target feature extraction, and data compression and encoding processes. The barrier issues in developing efficient detection, feature extraction and data compression/encoding algorithms for sparse distributed microphones are also discussed. Section 3 introduces

[1] OmniSence Specification sheet for Sensor Units Omni-410, 420, 430, 440 series, System Innovations Inc. http://www.system-innovations.com/

the proposed joint subband detection, feature extraction and data compression/encoding method and its constituent subsystems. These include: subband target attribute detector, subband-based feature extractor, data encoder and decoder, wideband DOA estimator, subband-based target classifier and finally coherent estimator for sparse arrays. Section 4 presents the test results of our developed algorithms on two data sets, namely the SAFE II and EAAGVS [2] acquired from the US Army-TACOM-ARDEC. The performance plots of the overall joint feature extraction-data compression system in terms of bit rate versus distortion, bit rate versus DOA error and finally distortion versus number of subbands are generated and thoroughly studied. We have investigated how the loss of information as a result of the data compression process would impact the quality of DOA estimation for multiple closely spaced targets and their classification into wheeled versus tracked types. Finally, Section 5 gives the concluding remarks and important topics for future research.

2 Identification of Barrier Issues & Review of Existing Methods

In this section, the main barrier issues in the development of efficient joint feature extraction, data compression, and encoding algorithms for a distributed array of single microphones for multiple target detection, classification, and tracking in realistic operating conditions will be investigated. Below we provide a review of some of the existing methods and present the most critical issues, based on the state-of-art knowledge in the literature, that are considered worthwhile for further investigation. It is our opinion that the issues indicated below present some of the most challenging directions to be researched in the next stages of this project and in the future.

2.1 Target Attribute Detection

Due to limitations in the bandwidth of the channel from each sensor to the master station, the data compression and encoding processes are too inefficient to be applied to the entire data recorded over a long period of time. Therefore, some type of a detection process is needed to screen the data for potential targets. In this case, only the data that appears to contain target-like acoustic signatures would be sent to the encoder. Furthermore, the detection algorithm should identify those frequency subbands that contain useful temporal-spectral information that is needed for accurate DOA estimation and target classification. An important consideration is the robustness of the detection process against wind noise, which interferes with the target signatures.

To develop reliable detection schemes for ground vehicles from their acoustic signals, it is essential to look at the temporal-spectral behavior of the target signatures. Observation of the spectrograms of the time series recorded by the microphone arrays in the real SAFE-II data indicated [6] that in cases when wind noise is not severe the time-frequency behavior of the targets exhibit peaks at certain frequency subbands. Clearly, this implies that the corresponding spectra of the time-windowed signals exhibit disjoint sharp peaks at frequencies where target indications are observed. To further validate this observation the time-frequency behavior of the acoustic signatures of ground targets in multiple target scenarios were also studied [6]. Again, it was observed that target attributes are present in disjoint narrow frequency subbands, hence validating the hypothesis that only a limited number of useful subbands may be detected at each snapshot or observation interval. The temporal-spectral features within the detected subbands may then be encoded and transmitted to the master station. It is expected that these features carry the necessary information for DOA estimation and classification of the sources. This will be demonstrated later in Section 4 of this final report.

[2] Enhanced Acoustic Algorithms for Ground Vehicle Surveillance, collected by SARA Inc.

Now, the question is that how does severe wind impact our detection and peak finding strategy? Again, our observation of the spectrogram of the SAFE II cases revealed [6] the fact that at time segments when the wind noise is severe and further there is no target indication, the power spectrum of wind noise exhibits decaying exponential behavior as a function of frequency. This property may be utilized together with a least squares (LS)-based fitting method to remove the effects of wind noise prior to target detection. This issue should be thoroughly studied in future research. Nonetheless, for low to moderate level of wind noise our proposed detection scheme in Section 3.1 performs very well without the need to remove the wind noise effects.

2.2 Time-Frequency Features & Data Compression

There are several related work [10]-[13] that use wavelet packets for time-frequency feature extraction of sonar and speech signals. In [10], the effectiveness of the wavelet packets was demonstrated on compression of speech with minimal distortion. It was observed that individual wavelet packets basis bear a striking resemblance to bursts of sound hence suggesting that these basis are well-suited for representations of acoustic and audio signals. In [11], Coifman and Wickerhauser used orthogonal wavelet packets basis to represent a variety of signals such as sound and images. They built a library of signal-dependent basis suited to the given signal or an ensemble of signals, which has the lowest information cost. Wavelets were constructed in [12] based on the acoustic wave equations to decompose and analyze complicated acoustic or seismic wave fields. Unlike the standard wavelets, propagation and scattering of acoustic wavelets become fairly simple. Comparison with standard discrete wavelet transform (DWT) revealed the efficiency of representing seismic data using acoustic wavelets. Two different subband encoding methods for compression of high quality audio signals are presented in [13]. These wavelet packets-based schemes use time-varying analysis and synthesis filter banks and block decomposition. The encoder and decoder use zero-tree algorithm originally developed for encoding of still imagery. The performance of the hybrid methods are shown to be superior to the MPEG Audio layer I standard. In [14], wavelet filters were applied to low bit rate audio encoding. Various codec models were designed and implemented using wavelet packets and an auditory perception model and entropy noiseless encoding. The method was shown to be better than the MPEG-audio layer I and II standards. A wavelet packets-based technique was used in [15] to perform feature extraction from the electromagnetic (EM) disturbance signal and classification of the extracted features in order to identify the possible causes of the disturbance. The same wavelet-based procedure was applied to signal compression and denoising as well.

More recently, there has been a growing interest in applications where speech acquisition is done using low cost, low power, and possibly mobile devices while the more complex speech recognition task is performed at a remote server [16]. This framework, which is similar to our problem, can be used either on the Internet or in wireless networks. An example is that a user employs a portable wireless device to access a remote speech-driven application. There are, however, only a few published papers [16]-[18] on the distributed recognition problem. Digalakis *et al.* [17] have shown that compressing the feature vectors that are used in speech recognition is indeed effective. They evaluated uniform and non-uniform scalar quantizers, vector and product-code quantization of the acoustic features and achieved bit rates between 2.6 kbps and 10.4 kbps. In [18], a new compression algorithm for encoding acoustic features used in speech recognition systems is proposed that uses a combination of linear prediction and multi-stage vector quantization. This compression algorithm can be used very effectively for speech recognition in network environments. It was shown that the computational complexity and memory requirements were modest for practical implementation. The algorithms were successfully tested on several test sets for several different languages demonstrating good performance with no significant change in the speech recognition accuracy due to compression. With this scheme they achieved a fixed rate of 4 kbps. In [19], the authors used Mel frequency Cepstral coefficients to represent speech features and hidden Markov models

(HMM) for recognition at remote station. They achieved bit rates lower than 1 kbps while providing good recognition. These recent studies reveal the fact that similar techniques could be potentially employed for data compression of acoustic signatures of multiple targets.

2.3 Encoding Methods

Once important target-useful subbands are detected, to represent and model spectral peaks or tonal frequencies in the selected subbands, the LPC method may be employed. The LPC scheme is widely used for speech encoding and recognition applications [20]. An N^{th} order linear autoregressive (AR) model that represents the m^{th} subband signal, $y_m(n)$, may be written as

$$y_m(n) = b_0\ e_m(n) - \sum_{i=1}^{N} a_i\ y_m(n-i), \tag{1}$$

where a_i's are the LPC or AR model coefficients and $e_m(n)$ is the driving process (or residual error), which is assumed to be white with zero mean and variance $\sigma^2_{e_m}$. The best N^{th} order linear minimum variance prediction of $y_m(n)$ is generated by $\hat{y}_m(n) = -\sum_{i=1}^{N} a_i\ y_m(n-i)$, where the predictor coefficients and $\sigma^2_{e_m}$ may easily be found using various parameterization methods [30]. Using the Yule-Walker method [30], these parameters may be found by solving

$$R_{yy}\mathbf{a} = \left[\sigma^2_{e_m}\ 0\ \cdots\ 0\right]^T \tag{2}$$

where $\mathbf{a} = [1\ a_1\ a_2 \cdots a_N]^T$ is the LPC coefficient vector, R_{yy} is the Toeplitz Hermitian auto-correlation matrix of the data in subband m. The linear predictor minimizes the variance of the residual error $e_m(n)$. This variance may easily be shown to be $\sigma^2_{e_m} = \mathbf{a}^T R_{yy}\mathbf{a}$.

A powerful benefit of the subband-LPC approach is that the same LPC coefficients can also be used to compress and encode the data. At the receiver, signal reconstruction can be accomplished using the decoded LPC coefficients together with the decoded residual error. This process is referred to as "predictive encoding" and has been found numerous applications in speech and image data compression. More specifically, the idea is to use the LPC model in each subband to generate an N^{th} order minimum variance linear prediction, $\hat{y}_m(n)$, of the data and then subsequently generate the residual error of the prediction, i.e. $e_m(n) = y_m(n) - \hat{y}_m(n)$ at each transmitter/sensor. The residual error and the LPC coefficients will be encoded and transmitted. The process of generating residual error removes mutual redundancy among successive samples. The residual error will have much lower dynamic range and hence can be quantized and encoded using substantially smaller number of bits.

At the receiver, assuming that the $(n-1)$ previously reconstructed values are available, we can generate $\hat{y}^o_m(n) = a^o_1 y^o_m(n-1) + a^o_2 y^o_m(n-2) + \cdots + a^o_N y^o_m(n-N)$, where superscript "$o$" implies the decoded/reconstructed value. The signal at time n is then reconstructed by adding the corresponding decoded error to the predicted value i.e.

$$y^o_m(n) = e^o_m(n) + \hat{y}^o_m(n)$$

The important observation is that $q_m(n) := y_m(n) - y^o_m(n) = e_m(n) - e^o_m(n)$ i.e. the quantization error of the predictive encoding (DPCM or differential PCM) is the same as that of the standard PCM. *However, for the same quantization error, $q_m(n)$, DPCM requires much fewer number of bits.* It can be shown that the improvement in bit rate for DPCM over PCM is given by $R_{PCM} - R_{DPCM} = \frac{1}{2}\log_2 \sigma^2_y/\sigma^2_e$ where R_{PCM} and R_{DPCM} are the rate distortion functions (bit per sample) of the PCM and DPCM, respectively. Clearly, $\sigma^2_y \gg \sigma^2_e$ which leads to $R_{DPCM} \ll R_{PCM}$, i.e. DPCM offers much better bit rate.

Obviously, the achieved compression ratio depends on how much redundancy is removed by the prediction. It must be pointed out that there are other redundancies in the data that can be exploited using the combination of the subband decomposition and LPC to achieve overall compression ratio of 50:1 or better. Oversampling the data contributes toward redundancy. For instance, for most of the targets the highest effective frequency does not exceed 250Hz. This implies that a large number of subbands do not contain any useful information. Additionally, the deployed microphones detect targets only in a small fraction of their operation cycle (about 90 days) in the field. Consequently, it is only necessary to detect and transmit data that contain militarily significant information. All such redundancies can be removed using the joint subband detection, feature extraction, and data compression schemes proposed in this project. Moreover, there is considerable amount of redundancy among recorded signals of multiple microphones that could be exploited to further reduce the amount of information, if the internal structure of a small subset of the microphones allows for both receive and transmit situations. In this case, for every small group or cluster of microphones there is an assigned 'leader' microphone that can receive the information from those microphones within the group, remove the redundancies and transmit only the novel information to the master station.

To achieve higher compression ratios and improve the robustness of the LPC method, one may use Cepstral representation [20]-[22] that is widely used for speech encoding and recognition applications. Using the Cepstral domain representation the dynamic range of the residual error is reduced even further, hence leading to much lower bit rate without sacrificing the accuracy. The Cepstral domain signal can be generated by taking the natural log operation of the magnitude spectrum (or power spectrum) and computing the inverse DFT of the result. An LPC model may then be fitted to the Cepstral signal, producing a residual error with much smaller dynamic range. To reduce the sensitivity of the high order Cepstral coefficients to noise, it is customary to window (e.g. Hamming window) the coefficients prior to LPC fitting.

Among the LPC coders, there are several primary candidates that can be applied. These are briefly reviewed below.

A. FS 1015 LPC

The FS1015 LPC [21],[22] was designed in 1984 for secure military applications. It can transmit speech at rates near 2.4 kbs; however, subjective speech quality assessments give it a poor rating. The nature of this project will not depend on subjective assessment of speech recreation, therefore some of the techniques associated with FS1015 may indeed be viable options for further investigation. FS1015 favors the use of covariance methods, rather than the autocorrelation methods which are more dominant nowadays. The FS1015 coder specifications were chosen with voice characteristics in mind. Mechanical generated acoustic sources do not share the wide variability of the human vocal system, hence some of the bit allocation for information associated with FS1015 (as well as other LPC methods) may be restructured to lend more accuracy where needed in the present project.

B. CELP

Code Excitation Linear Prediction (CELP) [21],[22] varies from the initial LPC coding method in that it uses multiple model excitation methods and an excitation codebook as well as a linear model codebook for passing the model parameters. The typical excitation for LPC models is a white Gaussian source. Although mechanical generated acoustic signal synthesis may benefit from research into source excitation, the current expectation is that the linear prediction models would be excited with the same source, be it Gaussian or some other type. CELP is mentioned here because it is a foundation of many of the standardized coders now in use. It is a medium bit rate coder, transmitting approximately 5 kbps, significantly more than that of the MELP coder.

C. MELP LPC

The Mixed Excitation Linear Prediction Coder (MELP) [22],[23] was designed in 1995 to overcome some of the limitations of other LPC coders while retaining the low transmission bit rate. Although it provides a subjectively high quality vocal reproduction, it maintains a bit rate near that of the FS 1015 coder. The term mixed excitation refers to the variability in speech requiring several methods of exciting the LPC model, depending on the nature of the speech sound to be reproduced. As indicated previously, mechanical generated acoustical signals do not have the same variability as the human voice, however, the need for multiple model excitation methods is not being ruled out.

D. GSM 06.10

GSM 06.10 [24] is the international standard digital mobile telephony encoding format. Although its 13 kbps transmission rate is higher than the FS1015 or MELP, it should be mentioned as the leading standard in mobile communications signal compression. GSM 6.10 utilizes regular pulse excited long term prediction (RTE-LTP) in its transmission scheme.

E. Non-Gaussian Excitation Functions

It has long been known in speech coding and synthesis that Gaussian excitation functions are insufficient for exciting or driving LPC models for speech synthesis. This problem has driven the generation of various excitation schemes such as RTE-LTP, MELP and CELP. The excitation schemes required for modelling the vocal tract do not bear much similarity to the expected excitation schemes for mechanical generated acoustical signals. However, it is noted that the detonation of the cylinders in a reciprocating engine acts similarly to a pulse train excitation of the vocal tract. Likewise, the exhaust system of a vehicle may be comparable to the vocal tract of a human being. Some applications using this concept may be of some use in modelling and encoding signals from engine driven vehicles.

F. Vector Quantization

Vector quantization allows for the minimization of the bits required for information transmission by assigning bit representations to values in areas with a high probability of occurrence. Because vector quantization does not send the exact value of the LPC coefficients, but rather reasonable approximations using a limited number of bits, there is always a loss of information involved. This error is minimized by assessing typical data and assigning bit representations centralized to the areas where the expected values are most likely to occur. Vector quantization is a feature of nearly all LPC coding schemes. The main concern with vector quantization is the efficient identification of the bit representation that most nearly describes the value to be transmitted. The listing of bit representations and the values they represent is referred to as a 'codebook'. The primary concern in vector quantization is the efficient choice of bit representations and the efficient navigation of the codebook once representations have been established [25]. The typical method of finding the correct representation is done as a binary-yes/no search. Some methods, such as those of [16], advocate the use of more than one codebook for obtaining good results. In the multi-codebook method, primary codebooks are used to quickly localize the area of the bit representation. Secondary codebooks are then employed to obtain the final representation, in a much faster fashion than using a yes/no tree search.

The main problem with the application of the LPC-based compression method for our specific acoustic signature compression application is that in each detected subband in addition to the model coefficients the residual error of the prediction must be encoded and transmitted. Although, the residual error has much lower dynamic range than the original signal and can be quantized and encoded using substantially smaller number of bits, the required bit rate of 1kbps for the available air-deployed acoustic sensors (e.g. OMNI-400 series) can never be achieved using this approach. Thus, taking into account the properties of source signatures and limitations of the available sensor hardware, it is required to design dedicated yet simple joint target detection, feature extraction, and data compression system for efficient sensor-to-

master station transmission. Using this method, only a small set of features that preserve important target attributes and time delays need to be encoded and transmitted. These features must be robust with no or minimal redundancy to yield maximum achievable compression rate.

In this Phase I research, a robust detection algorithm based upon the frequency spectra of the recorded signatures in small overlapping time windows is developed. This new spectral-based detection mechanism is employed at the sensor level to identify frequency subbands that could potentially contain target contributions. The idea is to exploit the peaky structure of the spectrogram of true sources and their possible harmonics in order to confine the range of frequencies for feature extraction, data compression, DOA estimation and classification to within those detected subbands. The motivations behind this idea include:

1. Only the features extracted from the data within the useful subbands in each snapshot need to be encoded and transmitted. This subband-based joint feature extraction and data compression method significantly reduces the amount of data that must be transmitted to the master station by removing the spectral-temporal redundancies in the data of each measured signal.

2. Noncoherent DOA estimation methods, such as the Geometric wideband Capon algorithm developed by ISTI [4], [26] can be applied more efficiently to those subbands that contain useful target features, hence eliminating the possibility of erroneous DOA estimation due to noise and wind effects outside the detected subbands.

3. The detection method can provide accurate DOA's even at far ranges as long as the corresponding signature can be detected in at least one subband in each snapshot.

4. The extracted subband spectral features in small time intervals could subsequently be used as clues to classify the sources. This idea is currently being utilized in most of speech recognition and speaker identification systems. Features extracted from additional snapshots can also be incorporated into the classifier to yield a high confidence classification decision.

5. The developed system is generic in nature using simple signal processing algorithms that are amenable for hardware implementation.

The developed joint feature extraction, data compression, and classification system and its constituent components are described in the next section.

3 A joint Subband Detection, Feature Extraction and Data Compression System

The developed system consists of several generic building blocks that are shown in Figure 1. These subsystems are described in detail in the following subsections.

3.1 Subband Target Attribute Detection

The main idea behind the subband detection approach is to detect the frequency peaks and the subbands that contain potential target information. The adopted frequency peak detection is based on the observation [6] that the frequency behavior of the ground targets of interest are not truly wideband and that target indications exist in disjoint narrow frequency subbands. Therefore, the spectrum of the recorded signal at each time instant has disjoint sharp peaks at frequencies where target indications are present.

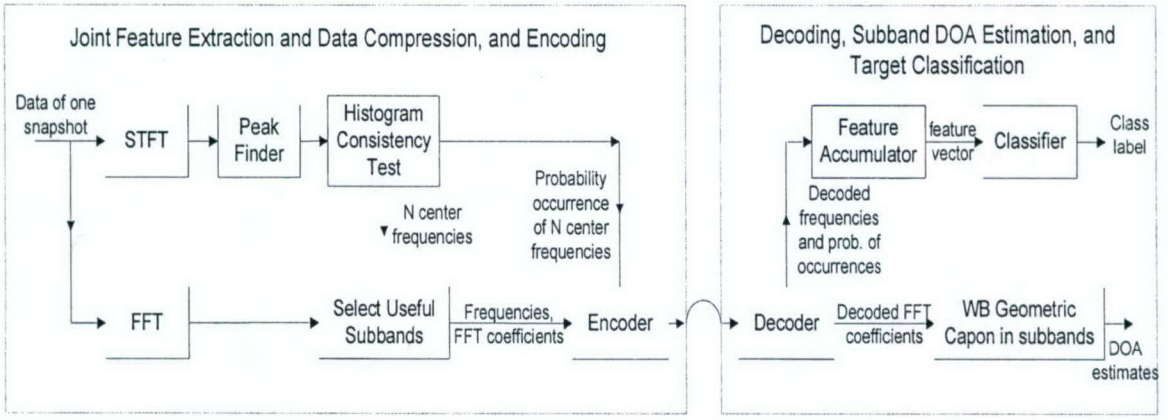


Figure 1: Block Diagram of the Proposed Joint Detection, Feature Extraction and Data Compression System.

On the other hand, wind interference has a relatively wideband frequency behavior which is exponentially decaying as frequency increases. Consequently, the presence of ground target may be easily detected by applying a peak finding algorithm to the spectrum of a signal recorded by a microphone at each time instant. This has been demonstrated in our first progress report [6] for this Phase I project. The steps of the developed detection algorithm may be summarized as follows:

1. Take the STFT of the recorded signal over a time-window. Use a window size of 256 samples within a window of 2048 points (two seconds of calibrated SAFE II data) with an overlap of 128 samples.
2. In each time-windowed spectrum obtained using the STFT, select up to N frequency peaks, in the frequency range of 31-250 Hz, that are above a chosen threshold. The frequency separation is 1 Hz.
3. Record the detected frequency peaks in the time-windowed spectra.

In Step 2, to detect the peaks of the time-windowed spectra, a peak finding process is used (see Figure 2). This scheme uses a sliding window (shown in dashed-line rectangle) to localize the maximum value of the time-frequency spectrum within the window. If the maximum happens at the center of the window and its value is greater than a 'threshold' the frequency that corresponds to this maximum is recorded as a peak point or a detection frequency. The window is then moved by 1 Hz and the procedure is repeated. The windows specified in red are those for which peaks at frequencies f_1 and f_2 are detected.

A simple way to determine a threshold for peak finding, which is somewhat adapted to the noise level in each sliding window procedure, is to find the median value of the portion of the spectra which lies inside the sliding window and then use a fixed pre-specified percentage (120%) of the median as the local threshold. We refer to this as "median-based thresholding". A potential peak is selected as a frequency peak if it is greater than this local threshold.

In the detection algorithm, the peak finding is employed for a time-windowed spectrum which represents the time-frequency behavior of the recorded signal over a very short period of time (1/8 of a second). In order to achieve accurate detection, it is important to determine if the frequency peaks selected for consecutive time-windowed spectra are indeed consistent. This can reduce the impact of having false peaks or a missing frequency peak at one or a few of the time-windowed spectra.

A simple way to perform a consistency test is to look at the histogram of the selected peaks over an observation period (2 seconds of calibrated data) and then select only those frequency bins that have a large

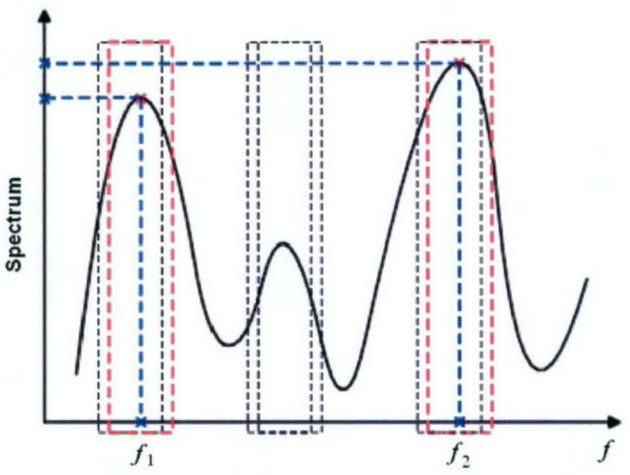


Figure 2: Peak finding using a sliding window. The detected peaks are located at frequencies f_1 and f_2.

number of occurrence by comparing them to a pre-specified threshold. In this way, only the most persistent frequency components are selected for further processing, e.g. feature extraction, data compression and DOA estimation. For the results in this research the threshold for selecting the frequency bins based on the histogram was set to 70% of the highest histogram bar. To reduce the computational load, the number of selected subbands has to be limited. Thus, we select only up to $N = 4$ frequency bins in every snapshot associated with the highest histogram bars that pass the threshold test.

3.2 Subband Feature Extraction, Encoding and Decoding Processes

After selecting the important peak frequencies using the above-mentioned detection scheme, we take the FFT of the calibrated signal of each microphone within 2 second time series. Note that in the calibrated data every 2 seconds interval corresponds to 1 second of the original recorded microphone data. Thus, one snapshot corresponds to 2 seconds of the calibrated data. Then, those FFT coefficients that correspond to the selected frequency peaks are chosen as center frequencies, around which frequency subbands of width 17 Hz (8Hz on each side) are built. The FFT coefficients that lie within these subbands are then selected as features that should be transmitted to the master station. This process is very similar in nature to wavelet packet decomposition [11] where the low energy subbands are eliminated. Consequently, as in the case of wavelet packet decomposition, it is rather easy to reconstruct, using inverse FFT, the time series of the recorded data using the retained frequency subbands, if needed.

In order to successfully transmit the extracted spectral features, an encoding scheme is required that allows for transmitting the features at a low bit rate without significant information loss. The encoding scheme we adopted here is very simple and is based on the IEEE-754 standard for conversion of floating point numbers to binary numbers and can easily be incorporated into any sensor communication system. Since the extracted features are spectral (FFT) coefficients, both magnitude and phase information need to be encoded and transmitted. It must be noted that even with this very simple encoding scheme the desired bit rate of 1kbps can easily be met. Clearly, using more elaborate encoding schemes such as Huffman or Rice encoding [8] substantially better bit rates can be achieved at the expense of increased system complexity.

The encoding scheme employed in this research may be explained as follows: Let a and ϕ be the

magnitude and phase of an FFT coefficient to be transmitted, respectively, and assume $-\pi < \phi \leq \pi$. We divide the magnitude a by 10^p, where p is the smallest integer for which $\hat{a} = a/10^p$ is smaller than 1. We then convert the pure floating point number $\hat{a} < 1$ from decimal floating point to binary using m_1 bits. We also convert the exponent number p from decimal to binary, using n_1 bits. Then, the binary codes for $\hat{a}$ and p are concatenated to obtain a code of length $m_1 + n_1$ bits, with the binary code for p being in the least significant bit positions of the code. This process completes the encoding of the magnitude a. At the receiver, the decoding process involves performing inverse operations to restore magnitude component a. For the phase component ϕ, we first divide ϕ by 10. This guarantees that $\hat{\phi} = \phi/10$ is a pure floating point number, i.e. $-1 < \hat{\phi} < 1$. We then convert $\hat{\phi}$ from decimal floating point to binary, using $m_2 + 1$ bits, where the most significant bit is reserved for the sign of ϕ. If ϕ is negative the sign bit is 1, otherwise is 0. This process completes the encoding of the phase component ϕ. Note that unlike what is done for the magnitude component, there is no need to code the power of 10 (i.e. 1) that ϕ is divided by, as we always divide ϕ by 10, even when ϕ is a pure floating point. Again, at the receiver the inverse of the encoding operations need to be performed to restore the phase information.

To further clarify the encoding process, here we provide a numerical example for encoding magnitude and phase of an FFT coefficient. Assume that the magnitude and phase of an FFT coefficient we want to transmit are $a = 2745.9682$ and $\phi = -1.1833$, respectively. The smallest exponent p for which $\hat{a} = a/10^p$ is a pure floating point is $p = 4$, which can be represented in binary using 3 bits, as 100. The decimal floating points number $\hat{a} = a/10^4 = 0.27459682$ may then be represented in binary using $m_1 = 9$ bits as 010001100. Concatenating, the binary codes for $\hat{a}$ and p yields the 12 bits binary code 010001100100 for the magnitude component a. For the phase component $\phi = -1.1833$, we divide ϕ by 10 to get $\hat{\phi} = -0.11833$. Then, we convert $\hat{\phi}$ to binary using $m_2 = 7$ bits, to get the binary code 10001111, where the first bit from left is the sign bit.

Our experimental results with the real SAFE II data revealed that the magnitude components in the frequency range of interest (31-250Hz) are indeed smaller than 10^4. Therefore, the largest value for the exponent p is 4, which can be encoded using 3 bits for this data set. Therefore, we always assign the first 3 bits for p. The number of bits (m_1, m_2) used to encode the pure floating point part $\hat{a} = a/10^p$ and that of the phase $\hat{\phi}$ determine the accuracy of the encoding process and the bit rate. Assuming that N_s essential frequency subbands are selected by the detection scheme, and there are 17 FFT coefficients in each subband, the bit rate required for transmitting the spectral features is

$$BR = \frac{17N_s(m_1 + m_2 + 4)}{1000} \; kbps \tag{3}$$

Note that this is the bit rate for a single microphone within an array.

Now, the key question is whether or not the essential properties of the original signals are kept in the decoded coefficients in order to perform accurate DOA estimation, target tracking and classification. In particular, it is important to study how choosing different number of subbands can impact the DOA estimation results and their accuracy in comparison with the true DOA's as well as the correct classification rate of the classifier. These, and other important issues will be thoroughly studied in Section 4.

3.3 High-Resolution Wideband DOA Estimation Method

The decoded FFT coefficients at the master station can be used to estimate the DOA's using one of the wideband Capon methods developed by ISTI [4],[26], namely Arithmetic, Geometric and Harmonic mean wideband Capon. These methods have produced [4],[26] great results for estimating the DOA's of multiple closely moving ground vehicles in the SAFE II data sets. It is interesting to note that we directly work with the decoded FFT coefficients without the need to reconstruct the time series. The reason being the DOA

estimation algorithm requires the FFT components and hence there is no point in going back to the time domain. Additionally, we shall show in Section 3.4 that the target classification can also be accomplished using these decoded FFT coefficients without reconstructing the signal in the time domain. However, if need be, this can easily be done by applying IFFT to the truncated frequency spectrum that contains only the non-zero components. This is a standard procedure which is also used in all of the transform domain compression schemes, such as discrete cosine transform (DCT), and wavelet transform-based compression [7]-[11] algorithms. In the following a brief description of the geometric wideband Capon algorithm, which is used throughout this research, is provided.

3.3.1 Geometric Wideband DOA Estimation

Consider an array of M sensors that receives the wavefield generated by d wideband sources in presence of an arbitrary noise wavefield that is assumed to be independent of the source signals. The array geometry can be arbitrary but known to the processor. The source signal vector $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_d(t)]^T$ is assumed to be zero mean and stationary over the observation interval T_0.

The array recorded signal vector $\mathbf{x}(t)$ is first decomposed into narrowband components by using a DFT over time segments of length ΔT. That is, the array output $\mathbf{x}(t)$, observed over T_0 seconds, is sectioned into K windows of duration ΔT seconds each. Thus, ΔT is the duration of one snapshot in the usual terminology of narrowband array processing and K is the total number of snapshots. We denote the jth narrowband component of all the outputs obtained from the kth snapshot by $\mathbf{x}_k(f_j)$, $k = 1, 2, \ldots, K$, and $j = 1, 2, \ldots, J$. The narrowband model for the recorded signal $\mathbf{x}_k(f_j)$ is

$$\mathbf{x}_k(f_j) = \mathrm{A}(f_j, \theta)\mathbf{s}_k(f_j) + \mathbf{n}_k(f_j) \tag{4}$$

where $\mathrm{A}(f_j, \theta) = [\mathbf{a}(f_j, \theta_1), \mathbf{a}(f_j, \theta_2), \ldots, \mathbf{a}(f_j, \theta_d)]$ is the $M \times d$ array manifold matrix of the sensor array system, with respect to some chosen reference point, $\mathbf{a}(f_j, \theta_i)$ is the steering vector of the ith source, $\mathbf{s}_k(f_j)$ and $\mathbf{n}_k(f_j)$ represent the jth narrowband component of actual source and noise signals, and $\theta = [\theta_1, \ldots, \theta_d]$ is the bearing angle vector of the sources. It is assumed that $M > d$ and that the rank of $\mathrm{A}(f_j, \theta)$ is equal to d for any frequency and angle of arrival. Further, it is also assumed that the decomposed narrowband components are independent.

Clearly, the goal is to determine the number of sources d and estimate the angles θ_i, $i = 1, 2, \ldots, d$ from the data $\mathbf{x}_k(f_j)$, $k = 1, 2, \ldots, K$; $j = 1, 2, \ldots, J$. For the narrowband case, the angle θ corresponding to the frequency component f_j may easily be determined using various algorithms [26], [27] such as Capon beamformer which minimizes the output power $q(f_j, \theta)$ with respect to θ:

$$q(f_j, \theta) = \frac{1}{\mathbf{a}^H(f_j, \theta)\mathrm{R}_{\mathbf{xx}}^{-1}(f_j)\mathbf{a}(f_j, \theta)} \tag{5}$$

where $\mathrm{R}_{\mathbf{xx}}(f_j) = \sum_{k=1}^{K} \mathbf{x}_k(f_j)\mathbf{x}_k^H(f_j)$ is the spatial covariance matrix at frequency f_j. The locations of the peaks of $q(f_j, \theta)$ correspond to the estimated DOA's.

For the wideband case, where all frequency components should be taken into account, the power at the output of the narrowband Capon beamformers, i.e. $q(f_j, \theta)$, may be averaged for all narrowband frequencies f_j, $j = 1, \ldots, J$. The locations of the peaks of the average output power determine the DOA estimates. Based on the averaging method used, several interesting wideband Capon DOA estimation algorithms may be developed, namely Arithmetic mean, Geometric mean, and Harmonic mean wideband Capon [26]. In the geometric mean wideband Capon, as evident from its name, the geometric average of the narrowband output powers is computed and used for wideband DOA estimation. That is, the angles θ are estimated by determining the locations of the peaks of the geometric mean power $Q_G(\theta)$:

$$Q_G(\theta) = \prod_{j=1}^{J} q(f_j, \theta) = \prod_{j=1}^{J} \frac{1}{\mathbf{a}^H(f_j, \theta)\mathrm{R}_{\mathbf{xx}}^{-1}(f_j)\mathbf{a}(f_j, \theta)} \tag{6}$$

Our experiment results indicated [4],[26] that the geometric wideband Capon algorithm usually provides more accurate DOA estimates than the other two wideband Capon methods. Thus, this method is used throughout this research.

3.4 Vehicle Classification Using Subband Features

To perform vehicle classification, specific spectral dependent features need to be extracted using the subband detection procedure. As mentioned before, the prominent subbands are selected based on the histogram consistency test. Unlike the DOA estimation process that relies on the decoded FFT coefficients in the detected subbands, our vehicle classification process requires the number of occurrences of the selected frequency peaks in 1 second each snapshot. Thus, the histogram consistency test not only provides N center frequencies but also their associated probabilities P_k's in every snapshot (2 seconds in calibrated data). This process is shown in Figure 1. These probabilities should also be encoded using the same encoding scheme. At the master station, the received features are decoded and accumulated over a period of 10 seconds in order to gather enough clues for accurate classification. The accumulated feature vector, which can potentially be of size 13 (i.e. covers 31-250Hz with 17Hz separation) is then used to classify the target into one of the four known vehicle classes, namely light-wheeled (LW), heavy-wheeled (HW), light-tracked (LT) and heavy-tracked (HT). Note that the window length of 10 seconds appears to be optimum as the decision about the class membership cannot be made in smaller size windows and enough clues need to be gathered before final decision making. The classifier is a multi-layer back-propagation network [28].

In Section 4.3, our joint subband detection, feature extraction and data compression system has been tested for vehicle classification based upon the extracted spectral-based features. The results attest to the fact that subband feature extraction and compression processes indeed preserve adequate acoustic harmonic structure of the sources in the frequency domain to allow for good classification performance. Although all the operations are performed in the frequency domain, the time domain aspects of the signatures are also retained through the use of IFFT.

3.5 Optimum Array Formation and Coherence Analysis

The fundamental requirement behind all sparse array processing methods is that signals received at different sparsely distributed microphones or sparse sub-arrays are coherent. The loss of signal coherence would have a dramatic impact on the performance of the sparse array processing algorithms. There are many factors that may influence signal coherence such as the distance between the distributed sensors, variations in environmental, terrain, and operating conditions, wind effects, and presence of natural or man-made obstacles. Therefore, it is of utmost importance to analyze the variations in signal coherence with respect to these factors. In addition, the analysis of coherence between multiple microphones or sub-arrays of microphones can be used to identify groups of microphones that exhibit high coherence, and hence carry common information. This can provide the opportunity to form dynamic time-space varying sensory arrays that can offer better localization and tracking performance in multi-target scenarios. This can be done with the aid of the multi-channel coherence test of Cochran and Gish [29], which is briefly discussed next.

3.5.1 Multiple-Channel Coherence Analysis

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ be the data vectors corresponding to M microphones (M data channels) of equal dimensions. It is shown in [29] that coherence or similarity of these vectors can be measured by

$$\gamma^2_{\mathbf{x},M} = 1 - \frac{g(\mathbf{x}_1,\ldots,\mathbf{x}_M)}{\|\mathbf{x}_1\|^2 \cdots \|\mathbf{x}_M\|^2}; \quad 0 \leq \gamma^2_{\mathbf{x},M} \leq 1 \tag{7}$$

where $g(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ is the determinant of the $M \times M$ Gram matrix,

$$\mathrm{G}_{\mathbf{x},M} = \begin{bmatrix} < \mathbf{x}_1, \mathbf{x}_1 > & \cdots & < \mathbf{x}_1, \mathbf{x}_M > \\ \vdots & & \vdots \\ < \mathbf{x}_M, \mathbf{x}_1 > & \cdots & < \mathbf{x}_M, \mathbf{x}_M > \end{bmatrix} \tag{8}$$

where (i,j)th element of $\mathrm{G}_{\mathbf{x},M}$ is the inner product of $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e. $\mathrm{G}_{\mathbf{x},M}(i,j) :=< \mathbf{x}_i, \mathbf{x}_j >= \mathbf{x}_i^H \mathbf{x}_j$ and superscript H denotes conjugate transpose. Also, $\|\mathbf{x}_i\|^2 =< \mathbf{x}_i, \mathbf{x}_i >= \mathbf{x}_i^H \mathbf{x}_i = \mathrm{G}_{\mathbf{x},M}(i,i)$. Thus, the generalized coherence estimate $\gamma^2_{\mathbf{x},M}$ may be rewritten as

$$\gamma^2_{\mathbf{x},M} = 1 - \frac{\det \mathrm{G}_{\mathbf{x},M}}{\prod_{i=1}^{M} \mathrm{G}_{\mathbf{x},M}(i,i)} \tag{9}$$

The value of $\gamma^2_{\mathbf{x},M}$ measures the similarity or coherence among all $\mathbf{x}_1, \ldots \mathbf{x}_M$. More specifically, if $\gamma^2_{\mathbf{x},M} = 1$, the data channels $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are linearly dependent or perfectly coherent; whereas $\gamma^2_{\mathbf{x},M} = 0$ implies that the data channels are linearly independent or incoherent. This test is applied to several cases in the EAAGVS database and the results are presented in Section 4.4.

4 Test Results & Observations

4.1 Brief Description of the Acquired Data Sets

Two acoustic signature databases were used in this study. The first data (SAFE II) was collected using three baseline wagon wheel-type pattern array of five identical elements (microphones) with 2ft radius at a sampling rate of $f_s = 1024$Hz. Table 1 shows the distances between the three array nodes. The acoustic signature database contains several runs for different single-file formations. The runs consist of acoustic signatures of one or more vehicles of the four typical classes mentioned before. For Study 1-3, we used the data of two SAFE II runs, namely runs 508 and 526. Run 508 contains six targets that move in three separate groups. The first group contains a single LW vehicle (BRDM), the second group is formed of three HT vehicles (T72's), and the third group includes two HW vehicles (ZIL's). Run 526 contains a single moving HW vehicle (ZIL). For Study 4, the SAFE II data of run 510 was used. This run contains one LT vehicle (BMP). Finally, for the target classification study in Section 4.4 we used several single target SAFE II runs, namely runs 514 and 515 that contain signatures of one moving HT vehicle (M60), and runs 526 and 527 that contain signatures of one moving HW vehicle (ZIL).

The second data set for sparse array processing, referred to as EAAGVS, was collected by SARA Inc. and contained the recording of a HT vehicle. The description provided in this section is based upon the documentation accompanying the data set and our interpretation of the information provided to ISTI. According to this documentation, the common format of the acquired data sets are as follows:

- A beginning time period of 30-60 seconds of ambient noise recording to capture the characteristics of the background, i.e. a windy day if it occurs.

- A time period of about 60 seconds where an 81 Hz pure tone is played by the speaker as a reference tone in the typical frequency range of an armored vehicle.
- A time period where the sound of a stationary M60 tank with an idle engine 1800rpm is played via the speaker.
- A time period where a sound clip of the same M60 tank moving away (should be toward) the microphone is played.
- A time period of ambient noise is recorded again at the end of the run to capture the characteristics of the background, in case the characteristics of the system has changed during the run (e.g. the wind has come up).

The stationary M60 data was recorded by a microphone placed 20m from the left side of the M60 with the engine running at about 1800rpm. Supplied documentation stated that the moving target signal is the same M60 tank *driving away from* the fixed microphone. Both of the M60 signals were sampled at sampling frequency of 1024Hz, which is the same as that used in the SAFE II data. All of the prerecorded sounds were played through a stationary speaker placed on the ground, about 72m (North) from the center microphone in the measuring array as shown in Figure 3. The target sound intensity was set at 105dB using a hand-held sound meter.

The sparse array data sets were collected with a circular wagon-wheel-type acoustic array of five elements, where the four microphones on the wheel were moved farther apart for each subsequent test and data collection (see Figure 3). Table 2 gives the details of the microphone spacing and data properties for each case. All data sets were acquired by playing a source signal (i.e. M60 tank) from a stationary speaker, and then recording with the microphone arrangement. The only moving entity is the vehicle in the sound recording played on the speaker. As can be seen from the last column of Table 2, the recorded time series have variable lengths. Additionally, the target recording scenarios are widely different. Thus, for the sake of consistency in our analysis and benchmarking, only those data sets that were collected in the same settings were used in this study. This subset includes recorded data sets for 1m, 5m, 10m, 20m, and 30m microphone separations.

Based upon the results generated on the EAAGVS database in [6] and later in this section, the following observations are made. First, the recorded data sets in EAAGVS database do not sufficiently and closely resemble signatures of real vehicles in order to test the usefulness of the algorithms. Additionally, due to the casing of the speaker the sound does not propagate uniformly hence causing possible discrepancies in the microphone recording intensity. Second, the sources must be at far-field in order to guarantee theoretical validity of most of the DOA estimation algorithms which rely on plane wave assumption. Clearly, this assumption is not valid for this data set. Third, a data set with moving sources will be more useful for testing the performance of different DOA estimation and tracking algorithms.

Table 1: Distances between the sensory nodes in SAFE II database.

	Node 1	Node 2	Node 3
Node 1	–	531.20	589.48
Node 2	531.20	–	584.86
Node 3	589.48	584.86	–

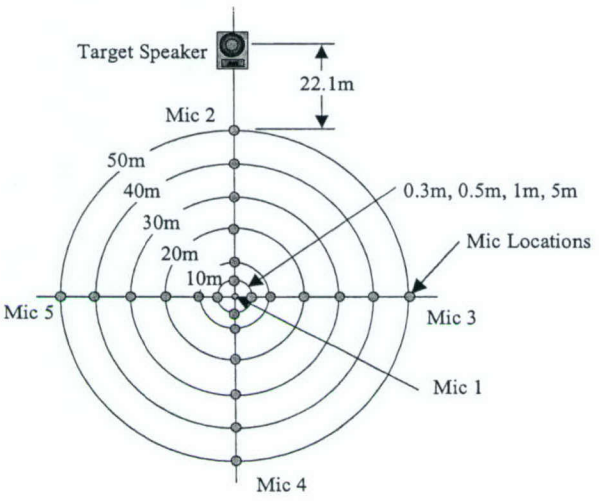


Figure 3: Sparse array data acquisition scheme and different array separations.

Table 2: Description of the data set characteristics for the EAAGVS data.

Data Set	Separation (m)	Description of Recorded Signal	Time (sec)
f4945	0.3	Ambient (Amb) Noise Alone	70
f5231	0.3	81 Hz Pure Tone Alone	75
f5848	0.3	Stationary (Stat) M60, Moving M60	117
f0808	1	Amb, 81Hz tone, Stat M60, Moving M60, Amb	297
f1707	2	Amb, 81Hz tone, Talking, Stat M60, Moving M60 (short), Amb	322
f2741	5	Amb, 81Hz tone, Stat M60, Moving M60, Amb	229
f3436	10	Amb, 81Hz tone, Stat M60, Moving M60, Amb	257
f4203	20	Amb, 81Hz tone, Stat M60, Moving M60, Amb	253
f4924	30	Amb, 81Hz tone, Stat M60, Moving M60, Amb	287
f5725	40	Amb, 81Hz tone, Stat M60, Moving M60, Amb	246
f0608	50	Amb, delay, 81Hz tone, delay, Stat M60, Moving M60, he'copter, Amb	543
f2008	50	Amb, 81Hz tone, End	202

4.2 Subband Detection, Feature Extraction and Data Compression Results

In this section, the algorithms developed for subband detection, feature extraction and data compression/encoding are tested and analyzed on the real SAFE II data sets acquired from the US-Army TACOM-ARDEC. These data sets involve different operating conditions and include single or several groups of multiple target scenarios in presence of additive interference. More specifically, we used the data of runs 526, 508, and 510 for this part of our study. The performance evaluation of the developed joint subband detection, feature extraction and data compression system is carried out using several measures that include: distortion in the retained signals, average bit rate and compression ratio, DOA error statistics, and robustness of the DOA estimation. In the next subsections, these results are presented for four different studies.

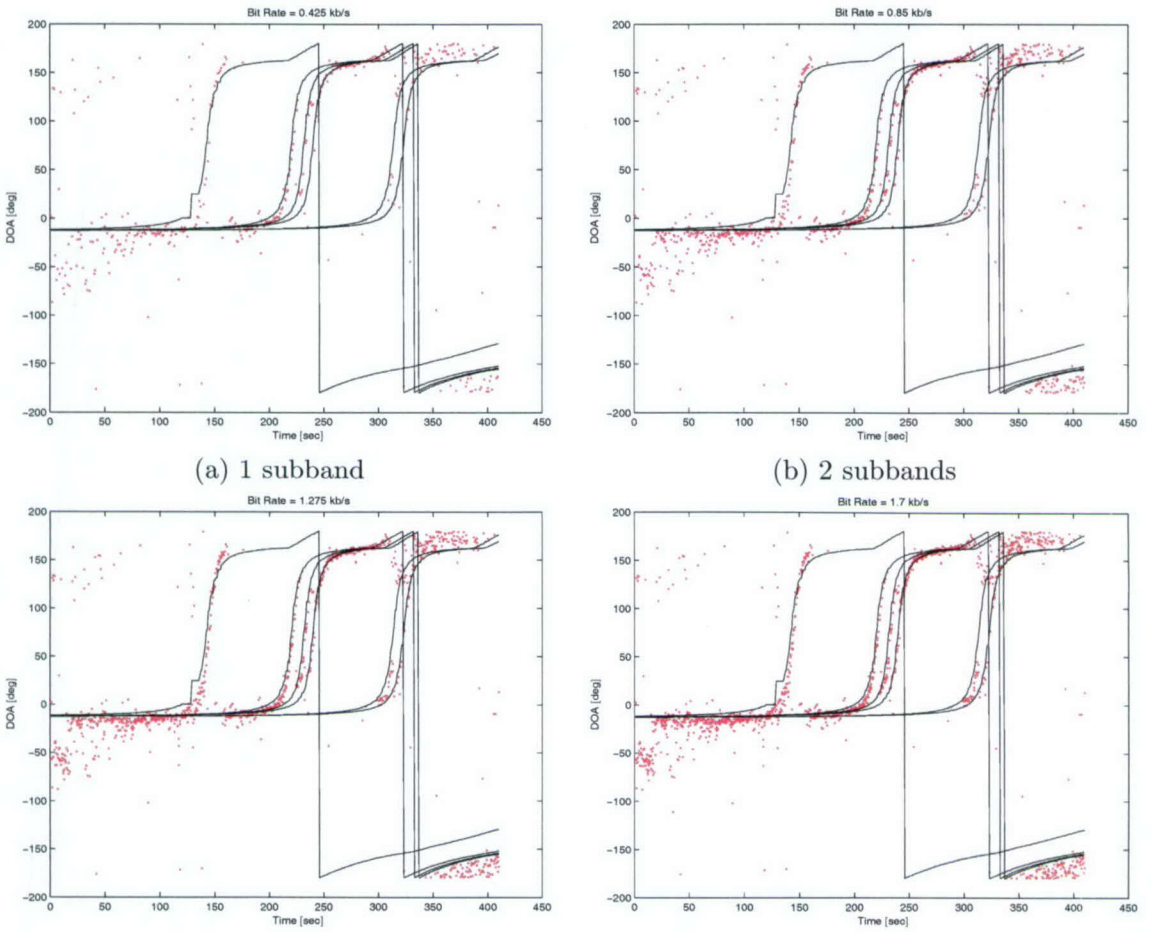
4.2.1 Study 1: Optimum Number of Subbands at a Fixed Bit Budget

In this study, the goal is to investigate the effects of the number of frequency subbands on DOA estimation accuracy given a fixed bit budget. Our aim is to study whether the information required for DOA estimation can be preserved in just a few essential frequency subbands, decided by the detection process. In the experiments in this section, the FFT coefficients are encoded with relatively high number of bits (15 bits for the magnitude and 10 bits for the phase) to ensure that the study will only reflect the effects of subbanding and not the encoding process.

The SAFE II data for runs 508 and 526 were used in this study. The first run, i.e. 508, contains the acoustic signatures of 6 moving targets, namely one BRDM, three T72's, and 2 ZIL's. The number of retained frequency subbands is varied from 4 to 1. In each case, the corresponding FFT features are selected, encoded, and then decoded, and subsequently used for DOA estimation using wideband Capon algorithm in [4], [26]. The feature extraction, encoding/decoding, and DOA estimation procedures described in the previous sections (see also Figure 1) were then applied. The DOA estimation in all the studies is performed using Geometric wideband Capon [4] algorithm.

Figures 4(a)-(d) show the DOA estimates for node 1 of run 508, obtained by retaining the FFT coefficients in one, two, three, and finally four frequency subbands, respectively. The results suggest that for this particular run going from one subband to two and then to three subbands does indeed improve the DOA estimation results. As can be seen in Figure 4(c), where 3 subbands were used, the DOA tracks are clearly identifiable and the targets are resolved at almost all snapshots. However, going from three to four subbands did not seem to improve the DOA results significantly. This implies that three frequency subbands is sufficient to capture the information required to estimate the DOA's and resolve the targets even in this multi-target run. This is very intriguing as we can localize the six targets by retaining only three frequency subbands, with 17 FFT coefficients in each subband. That is, only 51 carefully chosen FFT coefficients can preserve the information required for estimating the DOA's in run 508 of SAFE II data. The DOA estimates for nodes 2 and 3 of this run, obtained using different number of subbands are also shown in Figures 5 and 6, respectively. The same observations, can also be made for the results of nodes 2 and 3. These results demonstrate that the developed subbanding scheme offers very high data compression rate, without considerable loss of the essential DOA information.

It is also interesting to study a single target case, in order to see whether or not we can preserve the information required for DOA estimation by retaining only one frequency subband. Run 526 in SAFE II data set, which contains the signature of one moving ZIL vehicle, was used for this purpose. Figure 7 shows the DOA estimates for node 1 for this run, obtained when only one frequency subband is retained. As can be seen, the DOA's are correctly estimated at all snapshots suggesting that in this single-target case only one retained frequency subband is sufficient to allow DOA tracking of the target. Again, the results show



(a) 1 subband (b) 2 subbands

(c) 3 subbands (d) 4 subbands

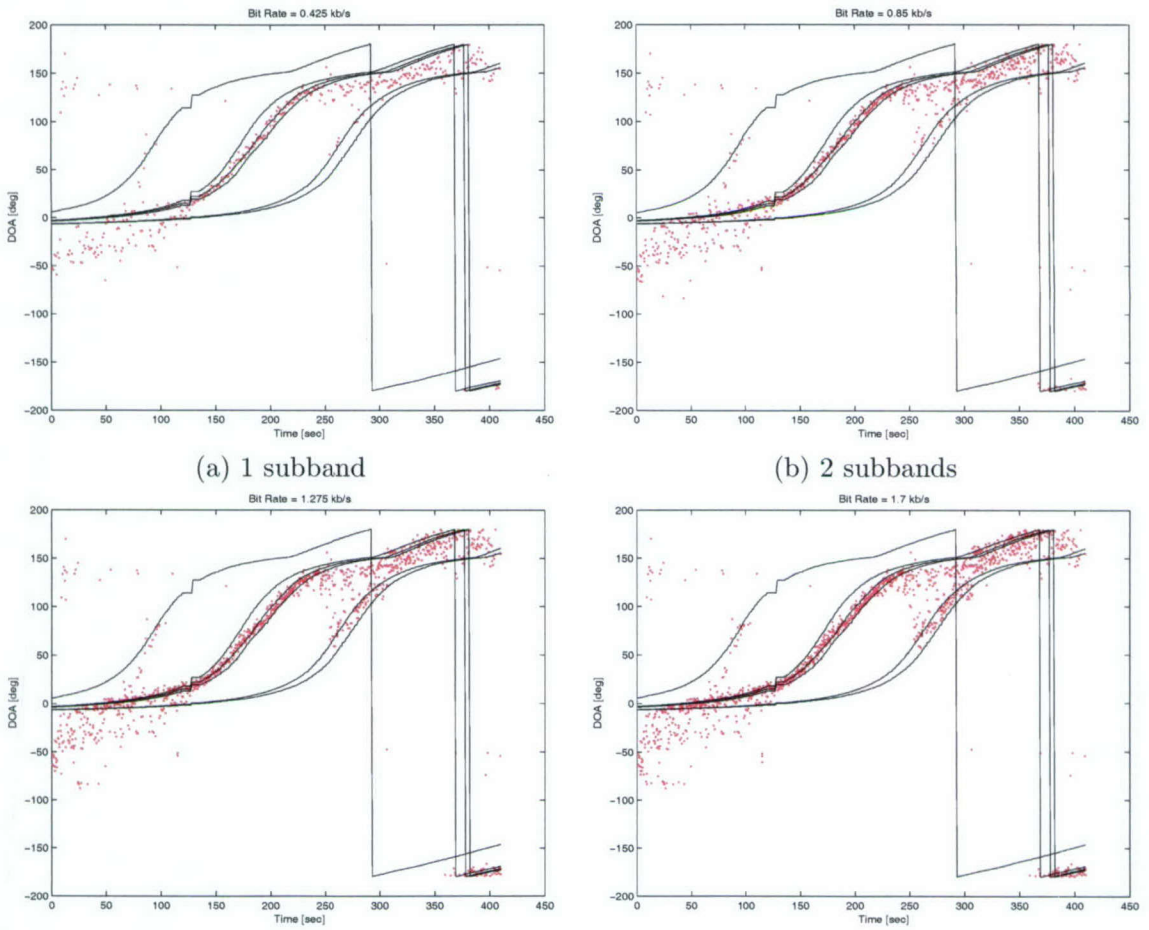
Figure 4: DOA estimates obtained by retaining different number of frequency subbands for Node 1 of Run 508 of SAFE II data; (a) 1 subband (b) 2 subbands (c) 3 subbands (d) 4 subbands.

that the subband decomposition method indeed results in a high compression rate without loss of useful DOA information. Note that the average bit rate per subband for each microphone is only 0.425kbps.

4.2.2 Study 2: Optimum Bit Rate for Fixed Number of Subbands

Now that we know how subbands capture the essential DOA information for a multiple-target scenario (e.g. run 508), we can turn our attention to the encoding/decoding processes. More specifically, we would like to determine: Given a fixed number of selected subbands at each snapshot, what is the minimum achievable bit rate to use for transmitting the corresponding FFT coefficients without loosing essential DOA information.

Again, the SAFE II data of run 508 is used here as in the previous study. As determined in Study 1, the optimum number of frequency subbands to be retained at each snapshot was 3 for this particular run. Therefore, at each snapshot $51 = 17 \times 3$ FFT coefficients have to be encoded and transmitted. In our experiments for this study, we vary the number of bits used in the encoding of the magnitude and phase components to generate different bit rates. The number of bits that are used here for encoding the



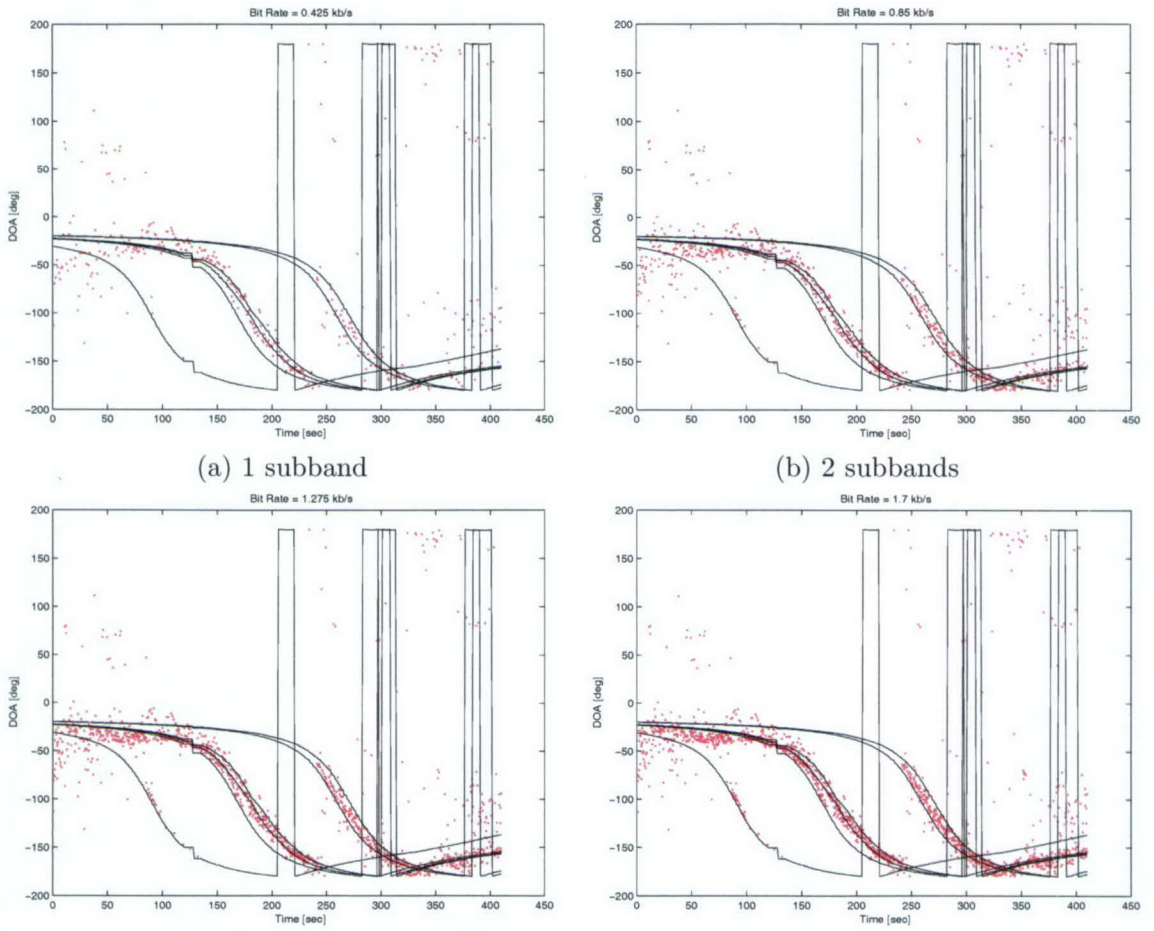
(a) 1 subband

(b) 2 subbands

(c) 3 subbands

(d) 4 subbands

Figure 5: DOA estimates obtained by retaining different number of frequency subbands for Node 2 of Run 508 of SAFE II data; (a) 1 subband (b) 2 subbands (c) 3 subbands (d) 4 subbands.



(a) 1 subband

(b) 2 subbands

(c) 3 subbands

(d) 4 subbands

Figure 6: DOA estimates obtained by retaining different number of frequency subbands for Node 3 of Run 508 of SAFE II data; (a) 1 subband (b) 2 subbands (c) 3 subbands (d) 4 subbands.

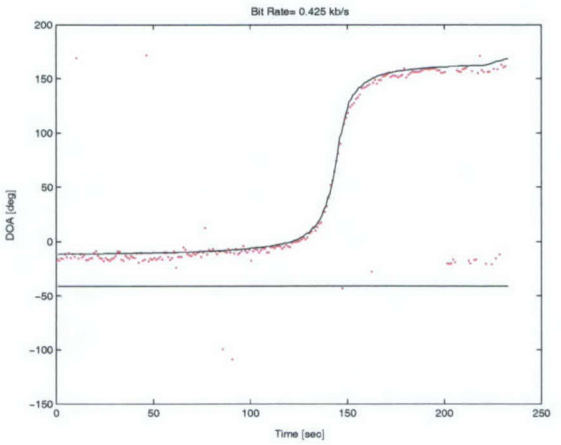


Figure 7: DOA estimates obtained by retaining only frequency subbands for Node 1 of Run 526 of SAFE II data.

Case No.	Magnitude	Phase	Bit rate [kbps]
1	19	13	1.632
2	15	10	1.275
3	11	7	0.918
4	7	4	0.561
5	5	2	0.357

Table 3: The number of bits used in encoding the magnitude and phase coefficients, and the average corresponding bit rates (one microphone and three subbands).

magnitude and phase components in different cases are given in Table 3, along with the corresponding average bit rate. Note that the bit rates are calculated per microphone, based on the fact that the FFT coefficients in 3 frequency subbands are transmitted at each snapshot.

In each case, the decoded FFT coefficients are used to estimate the DOA's using the Geometric wideband Capon [4]. Note that we directly work with the decoded FFT coefficients without the need to reconstruct the time signal. The reason being the DOA estimation algorithm requires the FFT components and hence there is no point in going back to the time domain. However, if needed, this can easily be done by applying IFFT to the truncated frequency spectrum that contains only 51 non-zero components.

Figures 8(a)-(e) show the DOA estimates for node 1 of run 508, obtained when the FFT features are encoded according to the bit rates in Table 3. As can be seen from Figure 8(d), the DOA estimates are very good even at bit rates as low as 0.561kbps for all the three subbands. In the first four plots, the DOA tracks are successfully identified and the targets are resolved at almost all snapshots. In the last plot, Figure 8(e), however, the DOA estimates do not follow the true ones, as the bit rate is too low (0.357kbps). Comparing the first four plots, it can be seen that using bit rates higher than 0.918kbps does not improve the DOA estimates noticeably. *Therefore, it appears that bit rate of 0.918kbps is the optimal choice.* This result suggests that it is possible to transmit the information required for estimating multiple DOA's at bit rates lower than 1kbps for multiple target runs and at the same time preserve the information needed for DOA estimation. As a result, this simple and generic joint feature extraction and compression algorithm can easily be incorporated into the existing single microphone sensors (e.g. OMNI-400 series) that have

simple off-the-shelf (OTS) DSP and communication system.

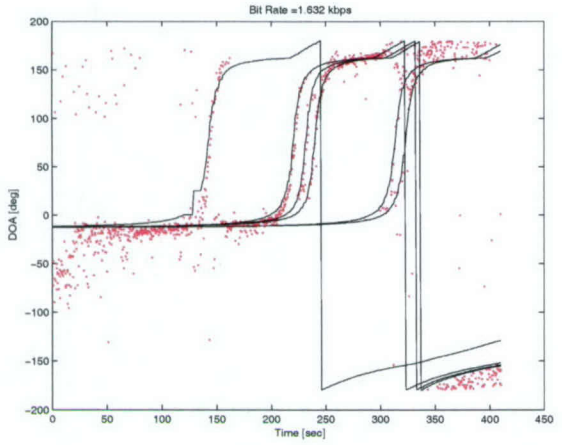
It must be emphasized that the bit rate achieved here is for three subbands. In a single target case, like run 526, where the data within only one subband needs to be transmitted, this rate can further be reduced to 0.306kbps. To demonstrate this, we carried out the same experiment with different bit rates on the data of run 526. Figures 9(a)-(e) show the corresponding DOA estimation plots generated using wideband Capon method [4]. The values of the mean and standard deviation of the absolute DOA errors (absolute value of the difference between the true DOA's and their estimates) at each bit rate are presented in Figure 10. The means and standard deviations, are computed based on the DOA errors at all the snapshots of run 526 (233 seconds long). It can be seen that the error statistics at bit rate of 0.306kbps are noticeably better than the error statistics at bit rate of 0.187kbps. However, the error statistics at bit rates of 0.425 and 0.544kbps are only slightly better than those at bit rate of 0.306kbps. *Therefore, this plot also suggests that the bit rate of 0.306kbps/subband offers the best overall performance.*

To further evaluate the performance of the encoding/decoding scheme, we may look at the noise-to-signal ratio (NSR) of the decoded FFT coefficients as a function of bit rate. This NSR is computed as the ratio of the energy of the difference between the original (before encoding) FFT coefficients and the decoded ones to the energy of the original FFT coefficient in the essential subbands. In order to obtain one NSR vs bit rate plot for the entire data of run 508 (412 seconds), the energies calculated at different snapshots were averaged. Figures 11(a),(b) show the plots of NSR vs bit rate for node 1 of run 508 of the SAFE II data, for the bit rates listed in Table 3 and for all the three selected subbands. Figure 11(a) is plotted in linear scale while Figure 11(b) is plotted in logarithmic scale. *These plots clearly show that the bit rate of 0.918kbps is at the knee point of the NSR vs bit rate curve. The NSR at this point is approximately -22dB (see Figure 11(b)), implying that the encoding is almost lossless.* Decreasing the bit rate beyond this point increases the NSR significantly. Therefore, these plots again attest to the fact that the bit rate of 0.918kbps (three subbands) offers the best overall performance.

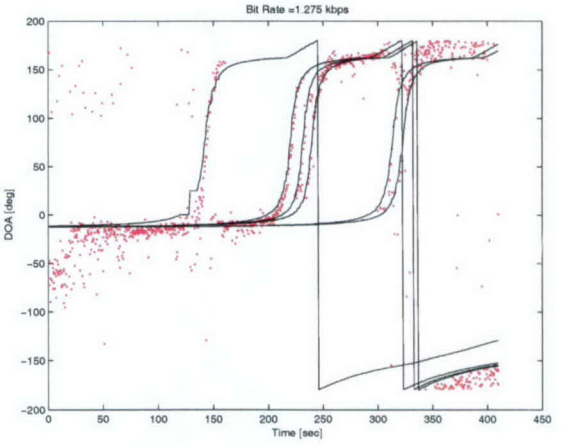
4.2.3 Study 3: Optimum Subband and Optimum Bit Rate

In order to further evaluate our feature extraction and encoding processes, in this section we compare the DOA estimates obtained based our subbanding scheme with those obtained based upon the entire frequency band (31-250Hz) without any compression and encoding. The purpose is to illustrate that the subbanding procedure followed by the encoding process does not deteriorate the quality of the DOA estimates considerably. Again, we used the SAFE II data of run 508, for which three essential subbands were retained as in Study 1. The corresponding FFT coefficients were then encoded at the optimal bit rate of 0.918kbps (for the three selected subbands) determined in Study 2. That is, the magnitude and phase coefficients are encoded using 11 and 7 bits, respectively.

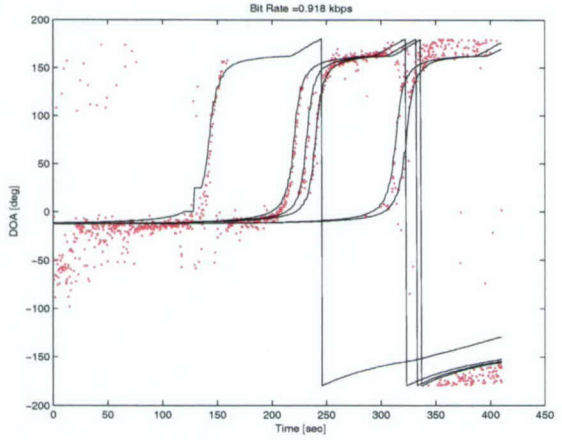
Figures 12(a)-(f) compare the DOA estimates obtained using the subbanding together with the encoding processes described before with those generated using the full frequency band without any data compression and encoding. Figures 12(a),(c) and (e) correspond to the case where subbanding and encoding are performed, while Figures 12(b),(d) and (f) correspond to the case where all frequency components are used without any compression/encoding. These plots clearly show that the subbanding and encoding processes using the optimal choices of the number of subbands (three) and bit rate (0.918kbps) do not deteriorate the DOA estimates considerably. Apart from some false DOA at the beginning of the run where targets are very far, the results obtained based on the proposed joint feature extraction and compression scheme are comparable with those obtained based on the entire frequency band without any compression and encoding. In fact, at some snapshots the results based on the proposed system are even better than those generated based on the full frequency band. For instance, as can be shown in the DOA plots for node 2 (Figures 12(c),(d)), the first target is detected at some snapshots, namely 90 to 100 sec, when



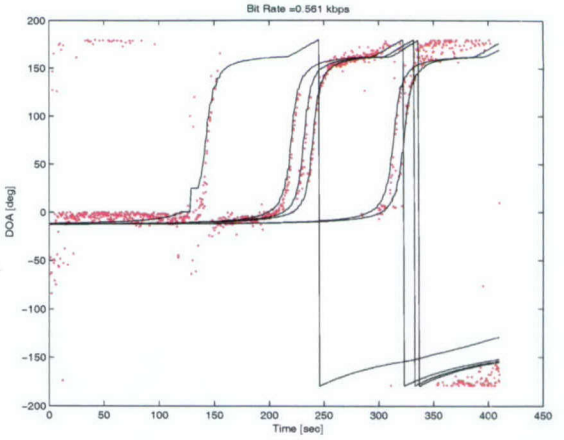
(a) Bit Rate = $1.63kbps$



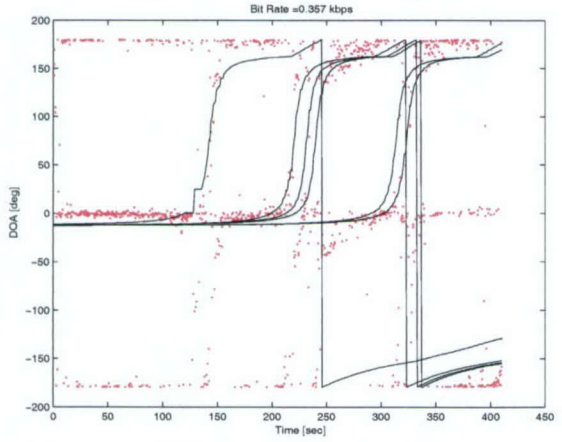
(b) Bit Rate = $1.275kbps$



(c) Bit Rate = $0.918kbps$

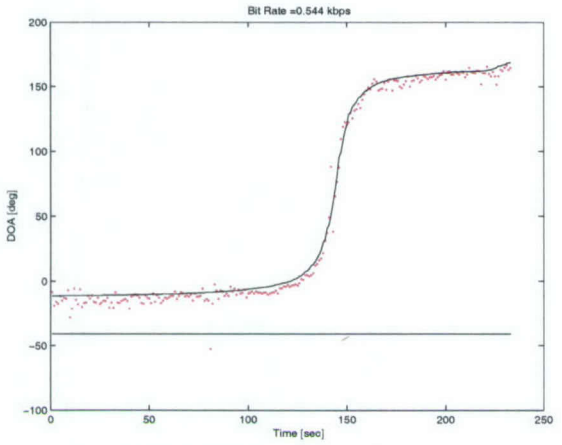


(d) Bit Rate = $0.561kbps$

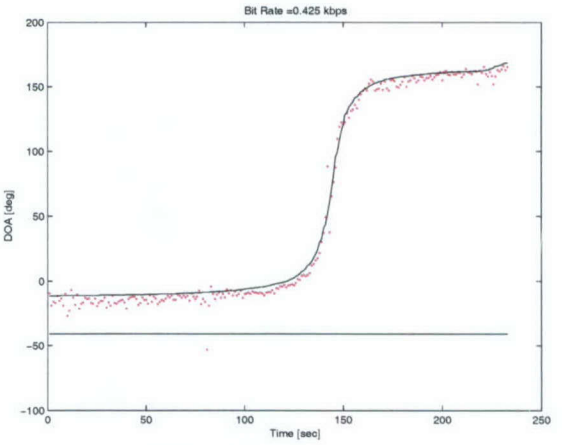


(e) Bit Rate = $0.357kbps$

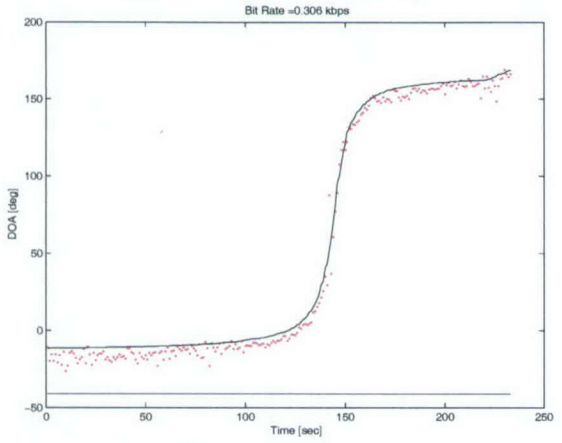
Figure 8: DOA estimates obtained by encoding 3 frequency subbands at different bit rates for Node 1 of Run 508 of SAFE II data; (a) Bit rate = $1.632kbps$ (b) Bit rate = $1.275kbps$ (c) Bit rate = $0.918kbps$ (d) Bit rate = $0.561kbps$ (e) Bit rate = $0.357kbps$.



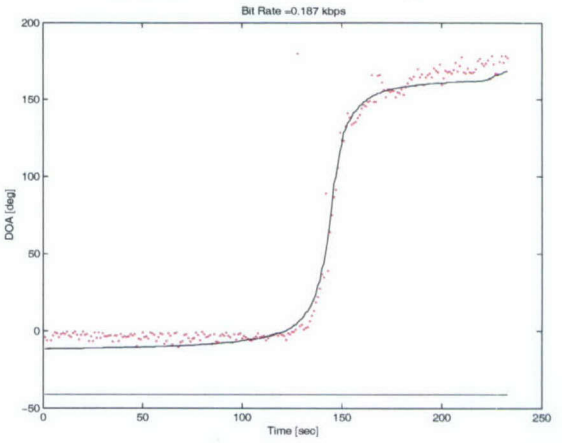
(a) Bit Rate = $0.544kbps$



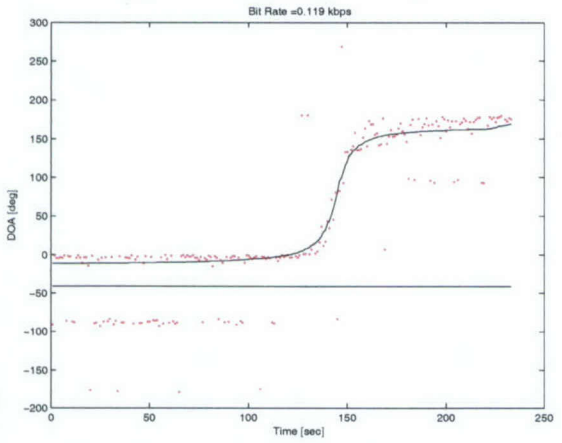
(b) Bit Rate = $0.4250kbps$



(c) Bit Rate = $0.3060kbps$



(d) Bit Rate = $0.1870kbps$



(e) Bit Rate = $0.1190kbps$

Figure 9: DOA estimates obtained by encoding 1 frequency subband at different bit rates for Node 1 of Run 526 of SAFE II data; (a) Bit rate = $0.5440kbps$ (b) Bit rate = $0.4250kbps$ (c) Bit rate = $0.3060kbps$ (d) Bit rate = $0.1870kbps$ (e) Bit rate = $0.1190kbps$.

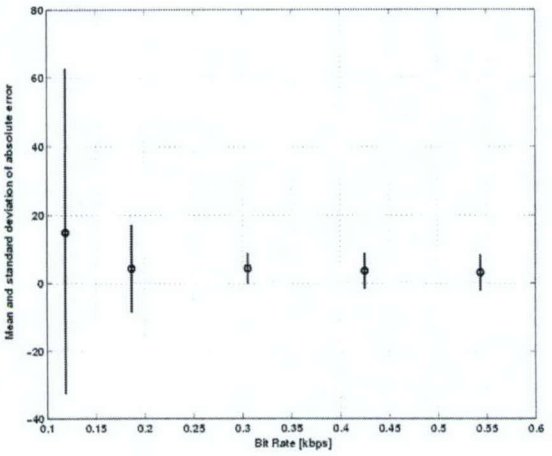
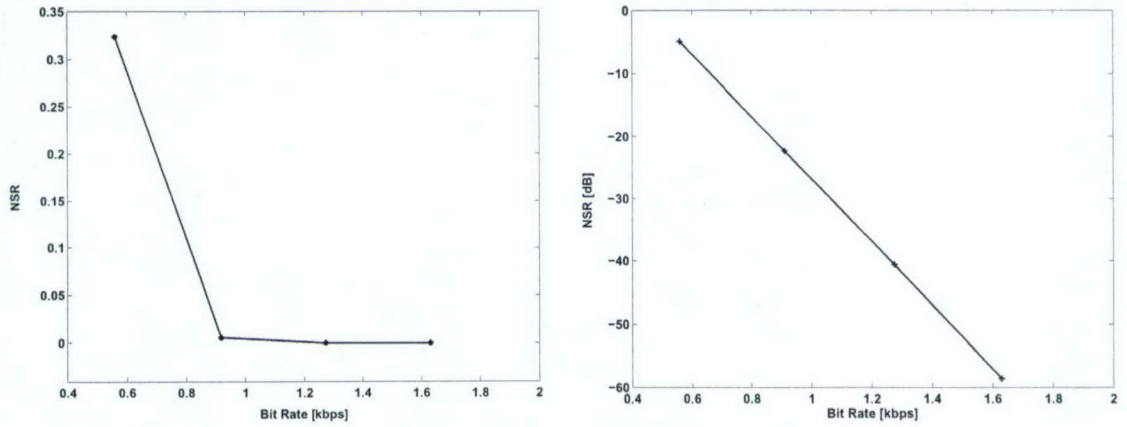


Figure 10: Plot of DOA error statistics vs bit rate for Node 1 of run 526.



(a) Linear Scale

(b) Logarithmic Scale

Figure 11: Plot of NSR vs Bit Rate for Node 1 of Run 508 of the SAFE II data.

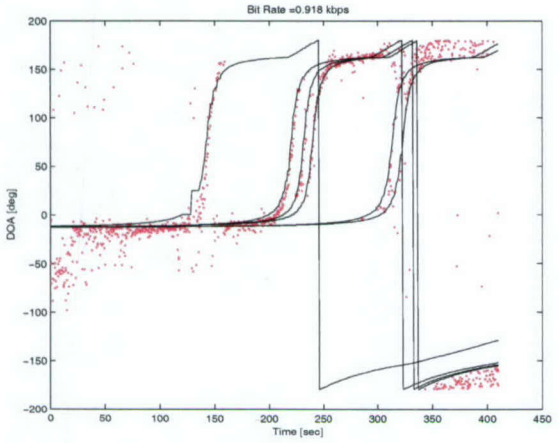
the proposed joint feature extraction and data compression scheme was used; whereas this target was not detected using the full-band DOA estimation. Also, it should be noted that subbanding reduces the computational load in the DOA estimation algorithm compared to the full-band case, as the number of frequency components that need to be processed is considerably smaller after the subbanding. This allows for faster decision making and fusion at the master station.

4.2.4 Study 4: Comparison with Wavelet-Based Method

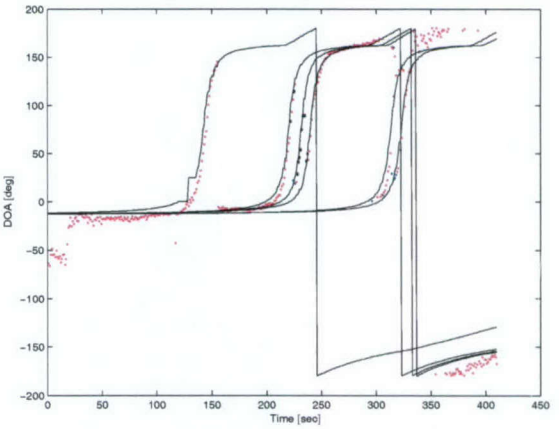
In this section, we provide a comparison between the developed joint FFT-based subband feature extraction and data compression method with the wavelet-based method described in Section 2.3. The comparison was performed on a 84 sec portion (133-216 sec) of data for node 1 of run 510 of the SAFE II data. This run contains one LT vehicle (BMP). Since this is a single target case, when applying the proposed subband feature extraction and data compression method of Section 3.2, only one subband needs to be encoded, and thus the bit rate will be 0.306kbps.

Using the wavelet-based method, the data of each sensory element (within the 84 sec time-window of interest) was partitioned into non-overlapping windows of duration 12 seconds each. This time duration was chosen to ensure capturing all the tones and spectral behavior of the target. A three-level wavelet packets decomposition (WP) [10] was then applied to the signal in each window. In WP decomposition, each subband extracts certain tonal features of the acoustic signatures. To avoid phase distortion that might be caused as result of using non-linear phase filters in the filter bank while at the same time ensuring the orthogonality of the signal representation, Symlet-4 wavelet [10] was used in the filter bank. The WP leads to 8 subband (filtered) signals, each with a bandwidth of 64 Hz and length of 1536 samples. Inspection of the frequency subbands shows that for the data of this study the frequency subband 64-128Hz contained most of the target attributes. Therefore, this is the only subband that was selected for encoding and transmission. An LPC model of order 12 was fitted to this subband in every 12 seconds time segment and the LPC coefficients together with the residual error were encoded and then transmitted. The other frequency subbands were zeroed out, and hence need not be transmitted. In the experiment conducted here, the LPC coefficients and residuals were encoded at a bit rate of 1.8kbps, using the same encoding method of Section 3.2. At the receiver the LPC coefficients and residuals were decoded, and the signal was reconstructed in the time domain. The wideband DOA estimation of Section 3.3.1 was subsequently applied to the reconstructed signal to determine the DOA estimates.

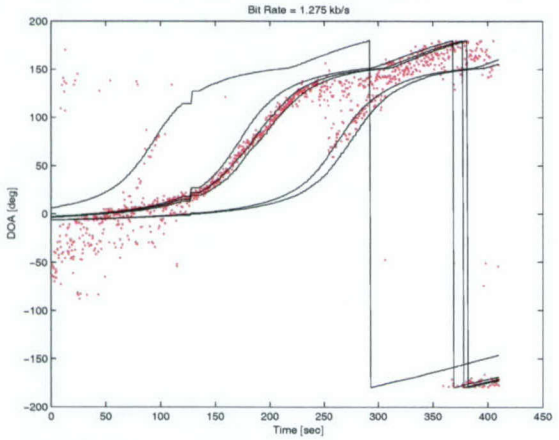
Figure 13 compares the DOA estimates obtained using the proposed FFT-based subband feature extraction and compression method (red dots) of Section 3.2, with those obtained using the wavelet-based method (blue crosses) of Section 2.3. The true DOA track is depicted by a black solid line to provide a reference. As can be observed from this result, the wavelet-based method even at a bit rate of 1.8 kbps does not provide accurate DOA estimates comparing to the results of the proposed method. The inaccuracy of the wavelet-based method could be attributed to the frequency leakage effects of the adjacent subbands. This is more prominent for short filters (e.g. Symlet-4) in the filter bank that have wider transition region in their frequency response. Additionally, taking into account the fact that the maximum achievable bit rate for each sensor is around 1.5 kbps, the results suggest that the wavelet-based method is not a good choice for the problem at hand. However, as can be seen the DOA estimates obtained using the developed FFT-based subband feature extraction and data compression method are accurate, even at a bit rate of 0.306 kbps. This clearly demonstrates the fact that the method developed in this Phase I research is a better choice than the wavelet-based method.



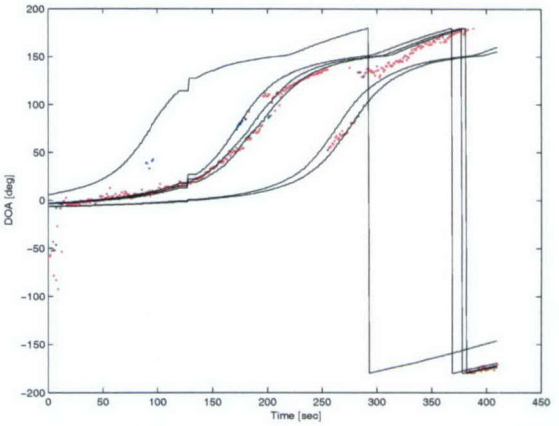
(a) Subbanding with encoding (Node 1)



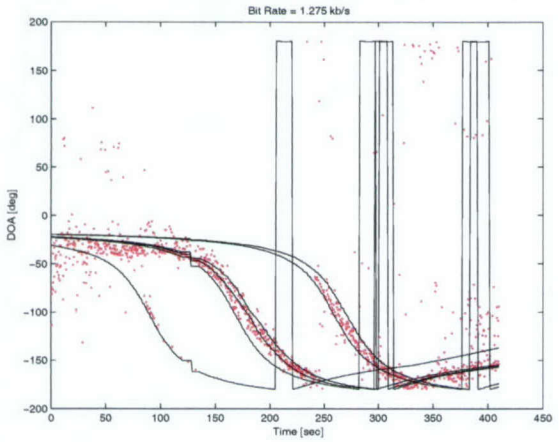
(b) Full-band without encoding (Node 1)



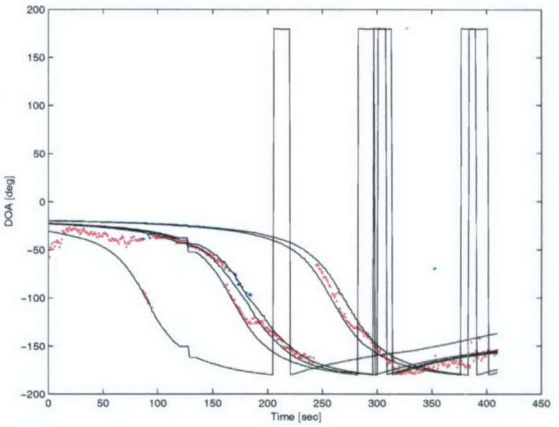
(c) Subbanding with encoding (Node 2)



(d) Full-band without encoding (Node 2)



(e) Subbanding with encoding (Node 3)



(f) Full-band without encoding (Node 3)

Figure 12: Comparison of the DOA estimates obtained using the frequency subbanding with encoding with those obtained using the full-frequency band and no encoding for run 508.

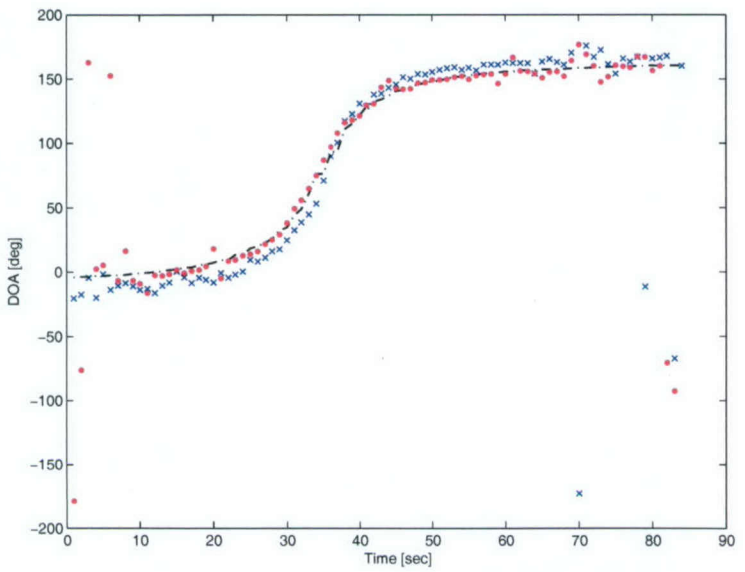


Figure 13: Comparison of the DOA estimates obtained using the FFT-based subband feature extraction/compression method with those obtained using the wavelet-based method.

4.3 Subband Vehicle Classification Results

The goal of this study is to evaluate whether or not the encoded information corresponding to only a few subbands is sufficient to perform accurate vehicle classification. For this preliminary study, a two-class classification problem is considered where the two classes are heavy tracked (HT) and heavy wheeled (HW). The training set consisted of spectral patterns extracted from the clean data of each of the classes. The clean signatures of M60 (HT) and ZIL (HT) vehicles provided by US Army TACOM-ARDEC were used for this purpose. The performance of the trained classifier was then evaluated on the real single-target SAFE II data for these targets.

As explained in Section 3.4, the subband based detection and features extraction process provides up to $N = 3$ frequency peaks and their associated probabilities in every 2 seconds of the calibrated data. However, for the training data since clean data is used no calibration was required. Thus, for the training data the frequency peaks and probabilities are generated for every 1 second time segment. The features extracted within every snapshot are the probabilities associated with the prominent frequency peaks in the 12 subbands (excluding the first subband). The decoded features for every 1 second are then accumulated over the 10 seconds observation period to form the final feature vector. It should be noted that, in the case of the training samples, the information related to all the 12 subbands is used in every 1 second time segment to form the combined 12-dimensional feature vector for the 10 second data. However, in the case of the testing samples, only the information related to selected (up to three) subbands in every 2 second (with 50% overlap) is used to form the composite feature vector. This is done intentionally in order to study the discriminatory ability of the decoded spectral features (testing) that represent only a small subset of all possible subband features (training).

Tables 4 and 5 show the clean data runs, the corresponding time segments where the spectral features were extracted (in every 10 seconds with 50% overlap) and the number of training samples extracted from each time segment. The M60 clean data led to a total of 100 samples, all of which were used for training of the classifier. For the ZIL data, although the total number of samples was 150, we only chose 20 samples from each run to give the same number of training samples.

Clean data	Time segment (sec)	# of Training samples
M60_RUN_893	40-120	16
M60_RUN_894	0-120	24
M60_RUN_895	40-120	16
M60_RUN_2089	0-100	20
M60_RUN_2090	0-120	24

Table 4: M60 clean data runs and time segments in which features were extracted for training the classifier.

Clean data	Time segment (sec)	# of Training samples
ZIL_RUN_182	0-100	20
ZIL_RUN_1454	50-300	50
ZIL_RUN_1455	30-250	40
ZIL_RUN_1456	30-130	20
ZIL_RUN_181	30-130	20

Table 5: ZIL clean data runs and time segments in which features were extracted for training the classifier.

The classifier was a simple three-layer back-propagation network (BPNN). Since this is a two-class problem, the number of output neurons of the BPNN is two. The number of neurons in the first and second hidden layers are 25 and 10, respectively. The transfer functions of the neurons in the hidden layers are hyperbolic tangential sigmoid and the transfer function of the output layer neurons are positive linear. The desired outputs of the classifier are $[1, 0]$ and $[0, 1]$ corresponding to HT (M60) and HW (ZIL) classes, respectively. The learning rate of the network was set to 0.05 and the maximum number of epochs was 2000. The classifier was trained exclusively on the features extracted from 10 second time segments of the clean signatures. The network was trained for 10 different random weight initializations and the network that gave the best performance was chosen. For this network, 100% correct classification rate was achieved on the training samples.

Testing samples were extracted from single-target SAFE II runs 514 and 515 that contain one moving M60 and runs 526 and 527 that contain one moving ZIL. In each of the specified runs, the calibrated data of the center microphones of nodes 1 were used. At the transmitter setup, up to 3 principal subbands were detected and the corresponding features were encoded and transmitted every 2 seconds. At the master station, the features received were decoded and then accumulated over a period of 20 seconds (calibrated) to form one accumulated feature vector. Again, note that every 2 seconds worth of calibrated data corresponds to 1 second of uncalibrated data. For each class, 15 testing samples were obtained from each run. This led to a total of 30 testing samples for each class. In the case of M60, 25 out of 30 samples were correctly classified i.e. about 83% correct classification rate; while for the ZIL 24 out of 30 samples were correctly classified leading to 80% correct classification rate. These preliminary results indicate that even with only 3 subbands, good classification performance can be achieved. In other words, essential information required for vehicle classification can still be retained within only a few selected frequency subbands determined by the proposed joint subband detection, feature extraction and data compression process.

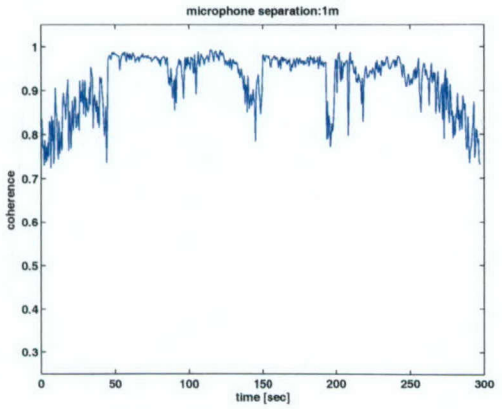
4.4 Coherence Analysis

In this section, the multi-channel coherence test in Section 3.5 is applied and tested on the EAAGVS data set, which was described in Section 4.1. The objective is to measure and analyze the coherence between the 5 microphones in the EAAGVS data with respect to different microphone configurations and spacing in order to determine whether or not sparse array processing methods are indeed applicable to the EAAGVS data set. The data vector for the i^{th} channel in the multiple-channel coherence test is formed of a block of the time series collected by the i^{th} microphones over 1/8 of a second. Since there are five microphones in the arrays, the multiple-channel coherence test measures the coherence among all five data channels.

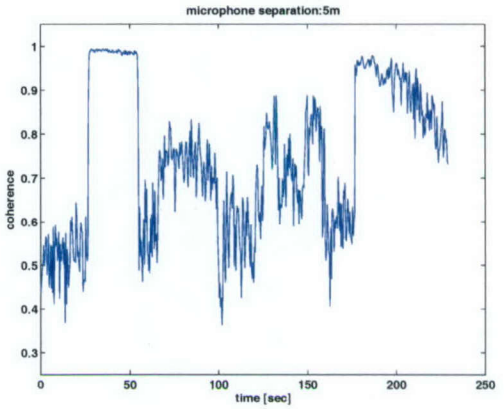
Here, we used the data of the runs for 1m, 5m, 10m, 20m and 30m sensor spacing in the EAAGVS database. All the conditions of the runs in terms of the type, number of sources, and their movement are the same. The coherence plots for these runs are shown in Figures 14(a)-(e) in the descending order of distance between the sensors. Overall comparison of all the coherence plots for different runs in Figure 14 together with the spectrograms in Figure 15 illustrate several interesting observations. First, at the time periods where the 81Hz reference tone is played, as the radius of the wagon-wheel array of microphones increases the coherence decreases. However, this decrease is very minimal due to the consistency and stationarity of this reference signal. The same conclusion cannot be drawn for the stationary and moving source (heavy tank) signals. This inconsistency in the coherence may be attributed to a number of factors related to this particular database. Therefore, one cannot necessarily associate this loss in coherence solely to the increase in sensor separation, though this could be one of the many reasons. To validate this claim, let us compare the coherence behavior of the stationary source in the 20m microphone separation case (Figure 14(c)) and the 30m case (Figure 14(e)) and their corresponding spectrograms in Figures 15(c) and (e). It is very intriguing to note that coherence measure for this stationary source is surprisingly much higher for the 30m case than for the 20m case, hence suggesting that there must be other factors that influenced the decrease in the coherence more dramatically than the increase in the microphone separation.

To show the usefulness of sparse array processing, the geometric wideband Capon algorithm in [4],[26] was applied to the EAAGVS data sets for various sparse array configurations with different sensor separation. In this case, the time series recorded by each microphone was partitioned into sliding windows of size 2048 (corresponding to 2 seconds of data) with 50% overlap. The DOA estimation results for the cases of 0.3m, 1m, 2m, 5m, 10m, 20m, and 30m microphone spacing are shown in Figures 16 and 17. In these experiments, the true bearing angle of the stationary source is around 180 degrees. Several interesting observations can be made from these results. As can be observed from the DOA results in Figures 16 and 17, microphone spacing of 0.3m provided less accurate but at the same time less ambiguous DOA estimates of the sources. As the microphone spacing increases from 1m to 30m, the DOA estimates become more precise. Nonetheless, more aliasing artifacts are also introduced which in turn increases the ambiguity in DOA estimation. The degree of DOA ambiguity in the estimates increases as the separation increases. Since the speed of sound in air is approximately 340m/s, the wavelength of the 81Hz tone is $\lambda \simeq 4\ m$. In the cases where the radius of the wagon-wheel array is larger than half-wavelength, i.e. $2m$, spatial aliasing and ambiguous DOA estimates are unavoidable for this particular frequency. Moreover, since essential spectral information of most of the sources of interest lie in the range of 50-250Hz the presence of aliasing is inevitable for all these sources when sparse arrays are used.

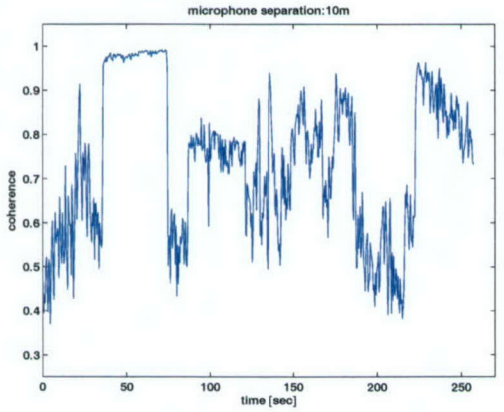
Comparisons of the coherence plots in Figure 14 with the corresponding DOA estimation results in Figures 16 and 17 clearly reveal that in situations where signal coherence was high, almost perfect DOA estimates were achieved. Even for nominal values of the coherence measure, DOA estimation are reasonably accurate. Although, from these figures it appears that these statements are valid for small microphone spacing ($\leq 2m$), the same observation holds true for larger spacing if one can resolve the ambiguity issues.



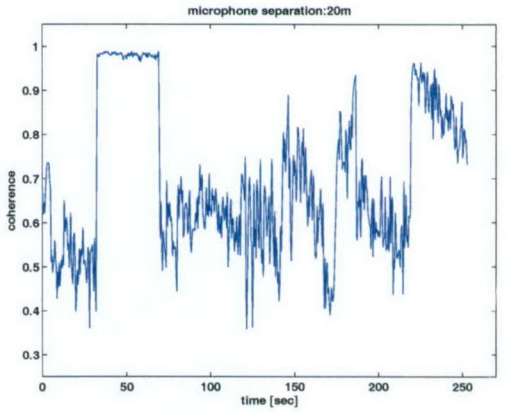
(a) 1m case



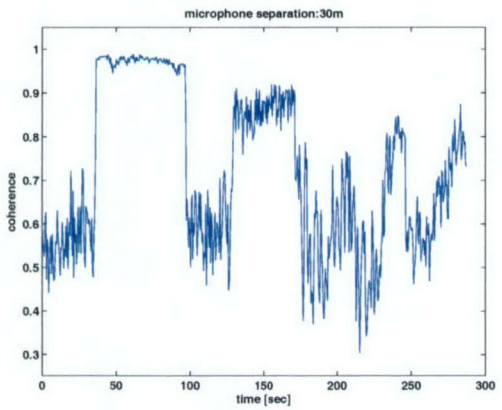
(b) 5m case



(c) 10m case

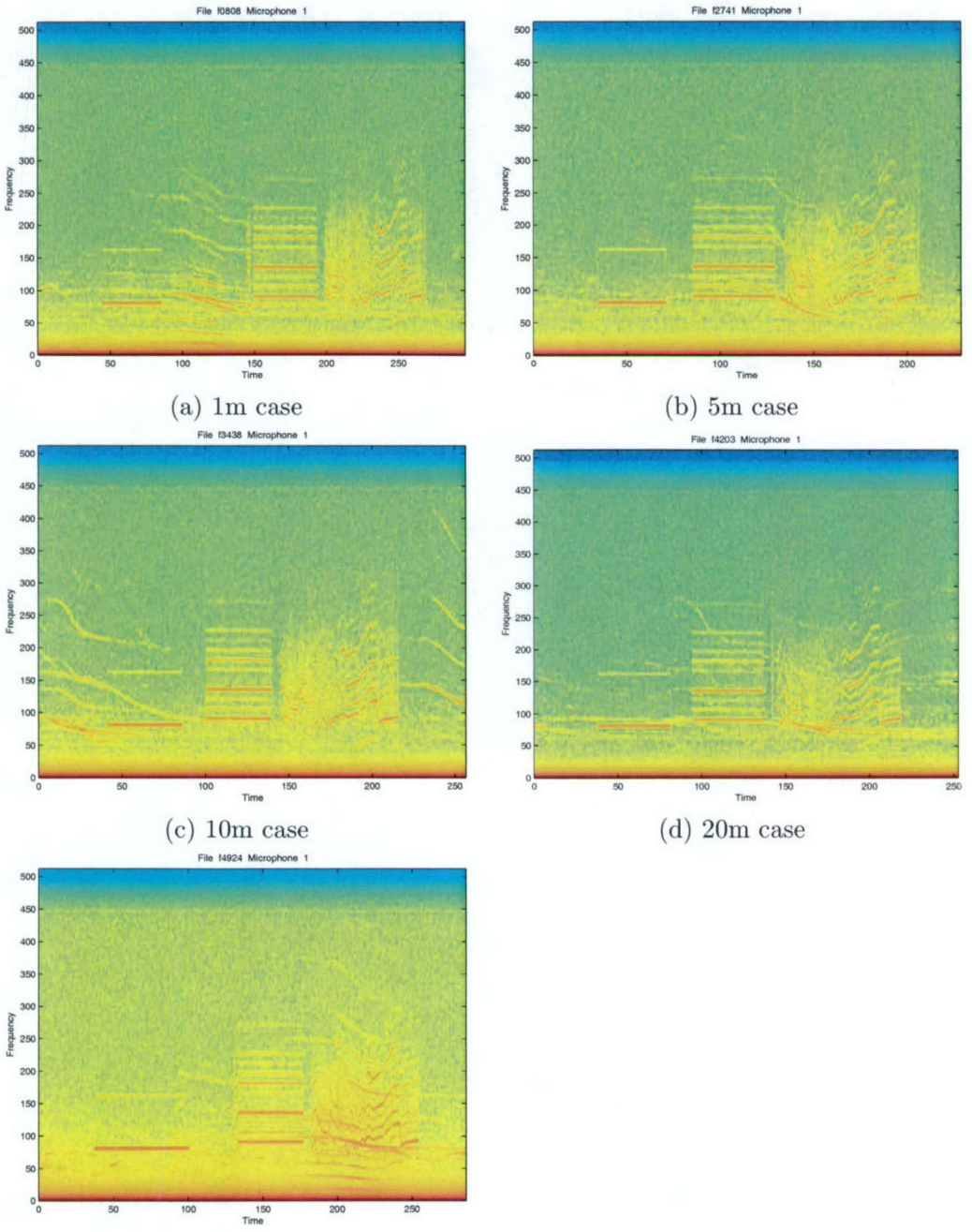


(d)20m case



(e) 30m case

Figure 14: Coherence plots for 1m, 5m, 10m, 20m, 30m microphone separation cases.



(a) 1m case

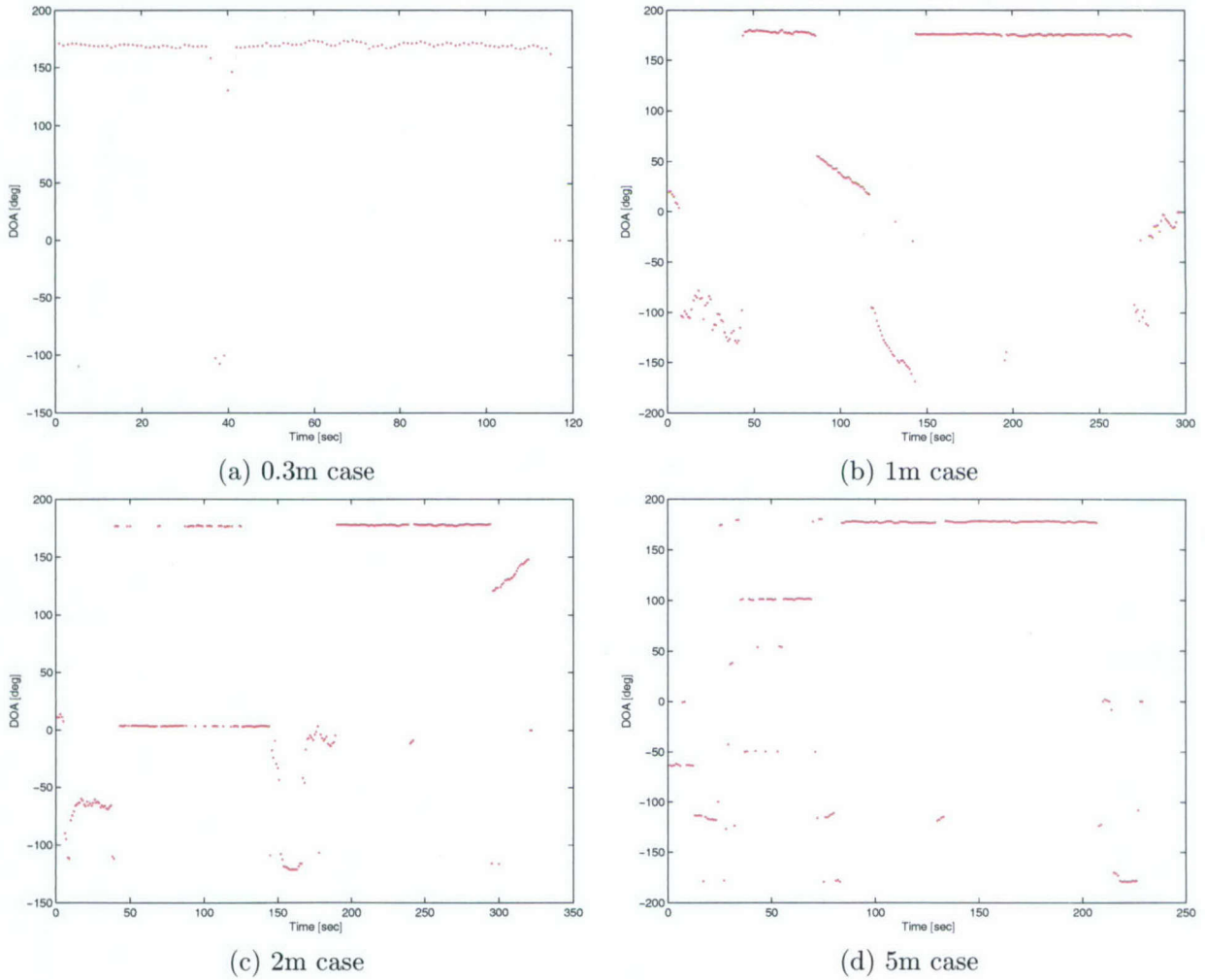
(b) 5m case

(c) 10m case

(d) 20m case

(e) 30m case

Figure 15: Spectrograms of the center microphones for 1m, 5m, 10m, 20m, 30m separation cases.



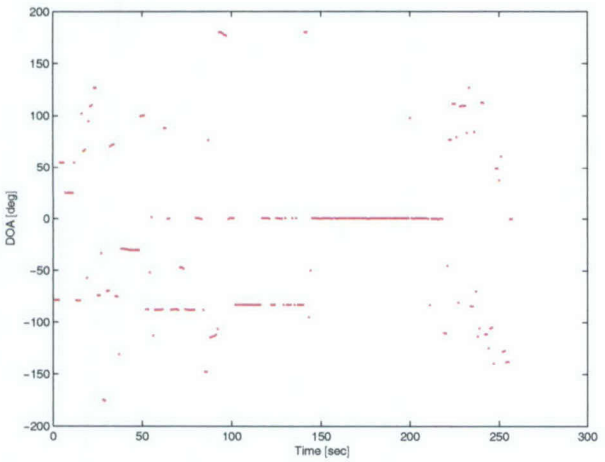
(a) 0.3m case

(b) 1m case

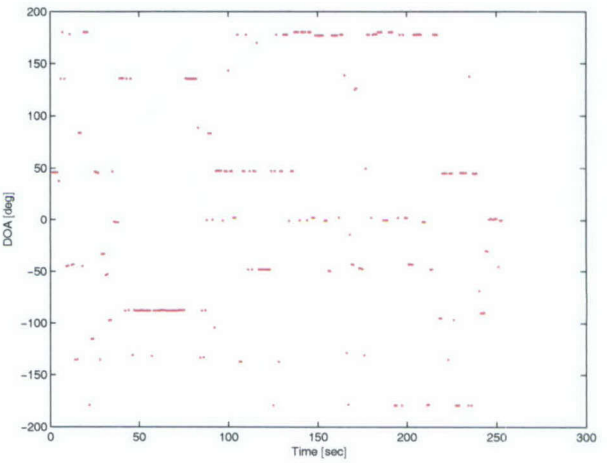
(c) 2m case

(d) 5m case

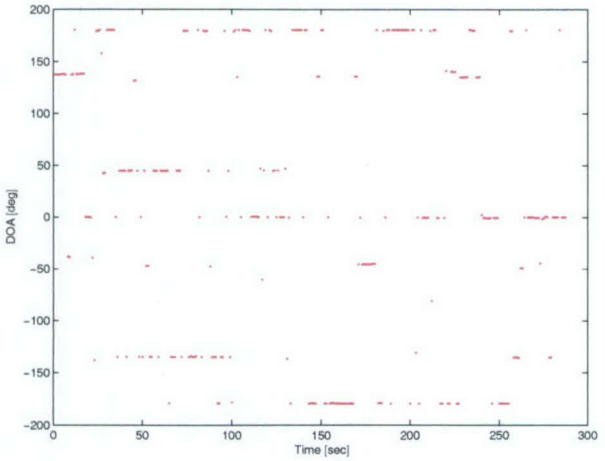
Figure 16: DOA estimates generated using wideband Capon.



(a) 10m case



(b) 20m case



(c) 30m case

Figure 17: DOA estimates generated using wideband Capon.

5 Conclusions and Suggestions for Future Work

The problem of detection, classification and localization of multiple ground targets, e.g. trucks and tanks, using several unattended single microphones is complicated due to various factors. These include: optimal deployment and placement of the isolated microphones in order to provide unambiguous target detection and localization, signal attenuation and loss of coherence as a function of range and doppler, effects of wind noise and weather/environmental conditions on the sensory data, compliance with the computational limitations of the processing boards, and effective communication of the detected features. As a result, the challenge in this problem is to develop new and yet simple algorithms to provide fast and extremely accurate detection of target attributes, efficient low bit rate encoding and quantization of the extracted features, and subsequent transmission to the master station. At the sensor level, however, developing such algorithms necessitates careful considerations for power consumption, size constraints, and complexity of the DSP and communication systems. At the master station, we require new algorithms that use the decoded features to determine the coherence measure of clusters or groups of microphones, define the optimal array or sub-arrays, perform wideband DOA estimation based upon the formed arrays or sub-arrays, combine the DOA's detected from the sub-arrays to resolve ambiguity and decide the final DOA's to accurately localize the sources, and finally classify the sources based upon the spectral-temporal behavior of the features over certain observation period.

In this Phase I research, we developed a new subband joint feature extraction and data compression scheme for multiple target detection and classification from acoustic signatures recorded using a distributed sparse array of several microphones. Various barrier issues and major impediments to the development of practical and real-life systems have also been identified. Based upon the results generated in this research, the following conclusions and suggestions can be given.

- By taking into account real operational constraints of the problem in hand, acceptable bit rate of the available air deployable acoustic sensors (e.g. OMNI-400 series), their on-board DSP capabilities and limitations, and algorithm complexity, we have developed a novel joint subband-based feature extraction and data compression scheme that preserves the essential features of multiple targets for subsequent DOA estimation at the master computer. The experimental results presented in this report indicated that even for multi-target runs an optimal bit rate of 0.918 kbps can be used to encode and transmit the essential spectral features within the detected subbands for each microphone. The decoded spectral features can be used in conjunction with frequency domain methods, such as wideband Capon [4],[26] to provide DOA estimates of the multiple targets without considerable loss of accuracy. The algorithms for detection, subband feature extraction, and data compression are very simple in structure and can easily be implemented using basic DSP boards for real operational settings.

- Clearly, it is possible to achieve higher compression ratios than what presented in this report using other transform-domain schemes such Discrete Cosine transform (DCT) instead of the simple FFT-based algorithm. Additionally, one could use more sophisticated encoding schemes such as Huffman, Rice encoding, or vector quantization [8] to encode the transformed coefficients using much smaller number of bits/coefficient. However, there is a trade-off between complexity of the data reduction and compression algorithms and the achievable bit rate. We believe our developed scheme is very simple, amenable for basic hardware implementation and at the same time provides excellent DOA estimation results even when compared with the results of full-band case without any data compression and encoding procedures.

- The proposed methods exploit subband spectral and tonal features of different (tracked or wheeled) target signatures which can be used not only for DOA estimation but also to provide indications

about the types (classes) of the targets, either prior to the encoding and transmission processes at the sensor level or after decoding at the master station. This is of great importance since it is desirable to perform pre-classification based upon some basic spectral features, e.g. center frequencies, at the sensor level. Then, depending on the target types and their military significance bit rate could be adjustable. However, more accurate target classification can be made based upon the decoded subband features at the master station without the need to convert the features to the time domain and an elaborate source separation process. Our preliminary classification results indicated that with only three selected subband features at every snapshot one can perform relatively accurate vehicle classification. The sequential hidden Markov model (HMM)-based classifier [31] developed by the PI can also be employed for determining the consistency of the features in several consecutive snapshots and performing high confidence "multi-look classification". This could be pursued in future research.

- We believe the results obtained in this report indeed show the promise of sparse array processing though there are numerous issues with the EAAGVS data set that must be resolved. As far as ambiguity problem is concerned, there are various schemes that can be employed to disambiguate the DOA estimates. The methods that exploits multiple invariance [32],[33] sub-arrays is certainly one typical approach that can be applied. Another method is to ensure that the array deployment takes advantage of various microphone spacing some of which are less than one-half wavelength of the source signal while the others are greater than one-half wavelength. In this way, for those groups of microphones that have spacing less than one-half wavelength of the source signal, coarse but unambiguous DOA estimates can be produced; whereas the sparsely located sub-arrays provide fine resolution for the DOA estimates at the cost of introducing ambiguity. Thus, by combining both properties, the unambiguous coarse DOA estimates may be used to eliminate the ambiguities introduced by aliasing.

- For the sparse array processing to work, we suggest conducting three different experiments using three different array configurations. In experiment 1, a three ring array (see Figure 18) is used. In this array configuration, the inner ring has eight elements, the middle ring has 6 elements and the outer ring has 4 elements. With this configuration, we have 4 circular array possibilities covering frequency subbands 246, 123, 92 and 41 Hz and 2 linear array possibilities that cover frequency bands 193 and 96 Hz. Thus, this configuration with 19 microphones gives great frequency diversity and it does not suffer from aliasing effects since each important subband has its own array. Also with this multi-ring array, we can get more accurate DOA estimation depending on range and frequency band. Moreover, this configuration does not have the problems with sparse arrays that exploit multiple invariance property, especially for near field DOA estimation when invariance is no longer valid.

  In experiment 2, random placement of microphones is considered. Here, we can either try to cluster the microphones in some randomly perturbed version of Configuration 1 (i.e. we still have the three clusters that offer frequency diversity but the structure is not regular), or randomly distribute them taking into account our clustering needs in terms of subband coverage.

  In experiment 3, four 5-element circular array (see Figure 19) may be employed. This experiment is only designed to test the multiple invariance properties. With this configuration, we don't have the frequency diversity that we have with the previous one. However, we can use the large aperture size for better resolution at the expense of ambiguity, which our proposed multiple invariance method is supposed to remove. The DOA estimates generated using these clusters of microphones can then be combined together in order to take advantage of the accuracy provided by the sparse arrays while removing the ambiguities introduced by spatial aliasing. Future research should address the problem

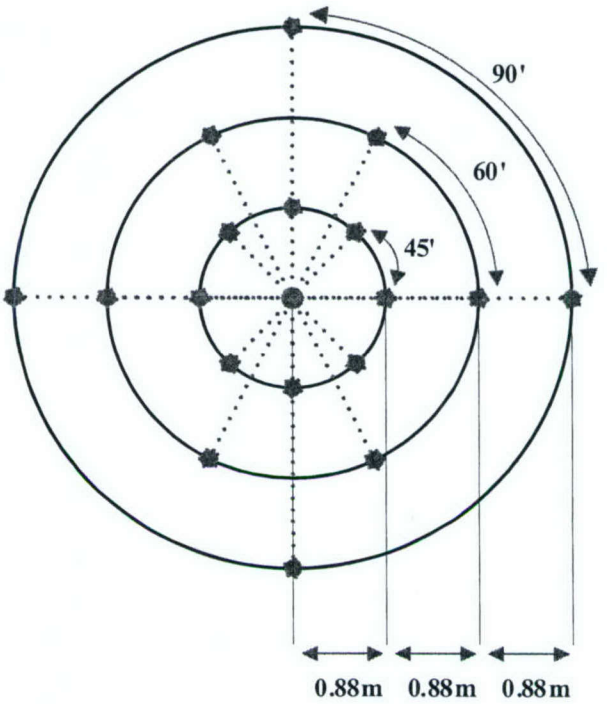


Figure 18: Three-Ring Circular Array Configuration 1.

of optimally combining different microphones in order to substantially improve spatial resolution for distinguishing multiple closely spaced targets in abreast, staggered or single-file formations.

- To identify clusters of microphones that possess good coherence for sparse array processing, the multi-channel coherence analysis tests can be applied at the master station. The multi-channel coherence test may also be used to refine the groups by either splitting or merging them into new groups that exhibit higher coherence. This method provides the opportunity to form dynamic time-space varying sensory arrays that can offer better localization and tracking performance in multi-formation and multi-target scenarios. The analysis of signal coherence can also be used to provide measures of performance in problems such as feature extraction, data compression, and encoding system. In other words, the success and effectiveness of a devised scheme for feature extraction, data compression, and encoding can be attested when the coherence among the group of the sensors in an array is preserved after the joint feature extraction-data compression process. These measures provide us with new additional tools to design an optimal acoustic signature data compression and encoding system.

References

[1] N. Srour, "Unattended Ground Sensors- A Prospective for Operational Needs and Requirements," *ARL Report Prepared for NATO*, October 1999.

[2] T. Pham and M. Fong, "Real-time implementation of MUSIC for wideband acoustic detection and tracking," *Proc. of SPIE AeroSense'97: Automatic Target Recognition VII*, Orlando, FL, April 1997.

[3] T. Pham and B. M. Sadler, "Wideband Array Processing Algorithms for Acoustic Tracking of Ground Vehicles," *ARL Technical Report*, Adelphi, MD, 1997.

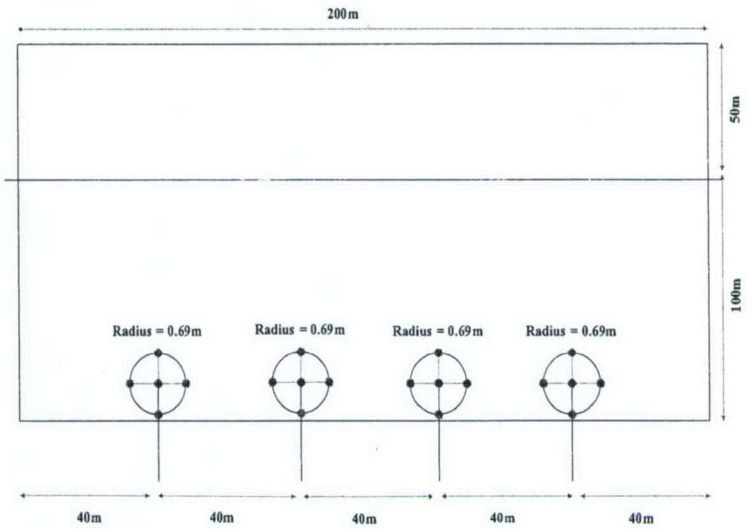


Figure 19: Four 5-Element Circular Array Configuration 3.

[4] M. R. Azimi-Sadjadi, "Detection, Tracking and Classification of Multiple Targets using Advanced Beamforming and Classification Methods: Phase II," *First-Year Summary Report, Phase II SBIR-Army,* January 2004.

[5] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate and capacity," *IEEE Trans. on Signal Processing,* vol. 48, 824-831, March 2000.

[6] M. R. Azimi-Sadjadi and A. Pezeshki, "A Joint Feature Extraction and Data Compression Method For Low Bit Rate Transmission In Distributed Acoustic Sensor Environments," *Progress Reports I,-III, Phase I SBIR-Army,*, 2004.

[7] N. Jayant, *Signal Compression: Coding of Speech, Audio, Image and Video*, World Scientific Pub Co., 1997.

[8] A. Moffat and A. Turpin, *Compression and Coding Algorithms*, Kluwer Academic Publishers, February 2002.

[9] S. J. Solari, *Digital Video and Audio Compression*, McGraw-Hill Professional, March 1997.

[10] M.V. Wickerhauser, "Acoustic Signal Compression with Wavelet Packets," in *Wavelets- A Tutorial in Theory and Application*, C. K. Chui, Academic Press, pp. 679-700, 1992.

[11] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. on Information Theory*, vol. 38, pp. 713-718, March 1992.

[12] R. S. Wu and J. H. Gao, "Application of acoustic wavelet transform to seismic data processing," SEG Expanded Abstracts, 1998 (Also in http://www.es.ucsc.edu/ wrs/publication).

[13] Y. Karelic and D. Malah, "Compression of high-quality audio signals using adaptive filterbanks and a zero-tree coder," *Proc. IEEE 18th Convention of Electrical and Electronics Engineers*, pp. 3.2.4/1-5, March 1995.

[14] P. Kudumakis and M. Sandler, "Wavelets versus conventional filters for low bit rate audio coding," *Proc. InternationalBroadcasting Convention*, pp. 320-324, September 1997.

[15] G. Antonini and A. Orlandi, "Wavelet packet-based EMI signal processing and source identification," *IEEE Trans. on Electromagnetic Compatibility*, Vol. 43, pp. 140-148, May 2001.

[16] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments," *Proc. IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 977-980, 1998.

[17] V. V. Digalakis, L. G. Neumeyer, and M. Perakakis "Quantization of Cepstral parameters for speech recognition over the World Wide Web," *IEEE Journal on Selected Areas in Communication*, Vol. 17, pp. 82-90 Jan. 1999.

[18] S. Bayer, "Embedding speech in web interfaces," *Proc. ICSLP*, pp. 1684-1688, Philadelphia, PA, October 1996.

[19] N. Srinivasamurthy, A. Ortega, Q. Zhu, and A. Alwan, "Towards efficient and scalable speech compression schemes for robust speech recognition applications," *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 249-252, July 2000.

[20] L. Rabiner, et. al., *Fundamentals of Speech Recognition.* Prentice-Hall, 1993.

[21] R. Goldberg and L. Riek, "A Practical Handbook of Speech Coders," CRC Press, Boca Raton, FL, 2000.

[22] W. C. Chu, "Speech Coding Algorithms," Wileys, 2003.

[23] D. Salomon, "Data Compression: The Complete Reference," pub-SV, 2004, 2nd edition.

[24] European Telecommunications Standards Institute, "ETS 300 580-2 Digital Cellular Telecommunications System (Phase 2)," European Telecommunications Standards Institute, 2000.

[25] Y. Linde and A. Buzo and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[26] M. R. Azimi-Sadjadi, A. Pezeshki, L. Scharf and M. Hohil"Wideband DOA Estimation Algorithms for Multiple Target Detection and Tracking Using Unattended Acoustic Sensors" *Proc. of SPIE-Defense and Security*, April 2004, Orlando, FL.

[27] H. L. Van Trees, *Optimum Array Processing,* Wiley Interscience, 2002.

[28] S. Haykin, *Neural Network: A Comprehensive Foundation*, Prentice-Hall, 2^{nd} Edition, 1999.

[29] D. Cochran, H. Gish, and D. Sinno, "A geometric approach to multiple channel signal detection," *IEEE Trans. on Signal Processing*, vol. 43, pp. 2049-2057, Sept. 1995.

[30] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forcasting*, Springer-Verlag, 2002.

[31] M. Robinson, M. R. Azimi-Sadjadi, and J. Salazar, "Multi-Aspect Discrimination of Underwater Mine-Like Object Objects using Hidden Markov Models", to appear *IEEE Trans. on Neural Networks.*

[32] K. T. Wong and M. D. Zoltowski, "Direction finding with sparse rectangular dual-size spatial invariance array," *IEEE Trans. on Aerospace Electrn. Syst.*, vol.34, pp. 1320-1327, October 1998.

[33] M. D. Zoltowski and K. T. Wong, "Closed-form eigenstructure-based direction finding using arbitrary but identical subarrays on a sparse uniform Cartesian array grid," *IEEE Trans. on Signal Processing*, vol. 48, pp. 2205-2210, Aug.